VOL. 8 No. 3, p. 65-96 MARCH 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, HOLLAND

BLOCKING-LAYER PHOTOCELLS

by W. CH. VAN GEEL.

621.383.5

Certain semi-conductors, such as copper oxide and selenium, in combination with a metal and separated from it by a very thin layer of insulating substance (blocking layer), exhibit upon illumination the so-called blocking-layer photoeffect. Upon illumination a photocurrent occurs without an auxiliary voltage being applied. In this article the action of such blocking-layer photocells and the mechanism of the phenomena occurring therein are discussed. In particular the selenium photocell is described in detail.

When light is allowed to fall on the alkali metal electrode of an evacuated tube or one filled with a rare gas, electrons are liberated. These electrons move in the surrounding gas or vacuum of such an alkali-metal photocell. By applying a suitable auxiliary voltage with respect to a metal anode, these electrons liberated by the so-called external photoelectric effect[1]) can be made to move toward the anode of the cell and thus give rise to a photoelectric current. In this way, however, only little electrical energy is obtained from the incident light, and it will usually be necessary to amplify the photocurrent generated before it is used.

In the case of the blocking-layer photocells to be discussed in this paper, however, the incident light is converted into electrical energy without it being necessary to apply an auxiliary voltage. The so-called blocking-layer photoeffect consists in the fact that the light causes electrons to move from a semi-conductor through the blocking-layer to the contiguous metal, from which they may then return to the semi-conductor through an external resistance (for example an ammeter). In this case, without further assistance a photocurrent of measurable size is immediately obtained, a fact which has of course very much promoted the employment of such blocking-layer photocells.

Although the blocking-layer photoeffect was discovered by Adam and Day in 1876, while it was again found by Fritts in 1884, blocking-layer photocells have actually only been developed within the last ten years to the state of technically useful instruments. Since then they have found extensive use in the measurement of intensities of illuminations as applied in photographic exposure meters, luxmeters, photometers, colorimeters and the like.

Structure of the blocking-layer photocell

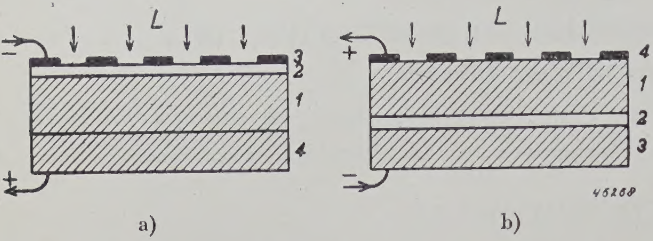
The construction of a blocking-layer photocell is in principle similar to that of a blocking-layer rectifier[2]). In both cases we are concerned with a metal layer and a semi-conductor which are separated by a thin insulating layer, the so-called blocking layer. In order to obtain a blocking-layer photocell, it is now necessary to make the counter electrode or the metal layer which makes contact with the semi-conductor so thin that it is transparent. The light can then penetrate into the semi-conductor and there free electrons.

Two kinds of blocking-layer cells can be distinguished according as the blocking-layer lies in front of or behind the semi-conducting layer and the electrons leave the side of the semi-conductor upon which the light is incident (front-wall cell) or the other side of the semi-conductor (back-wall cell). In *fig. 1* the principles of these two possibilities are shown diagrammatically. In this figure *1* is the semi-conductor and *2* the blocking-layer, while *3* is the metal counter electrode toward which the photoelectrons from the semi-conductor move through the blocking-layer. The layer of metal in contact with the semi-conductor, which serves the purpose of causing the return of the electrons to the semi-conductor after traversing the circuit, is indicated by *4*. The light is incident in the direction of the arrows, so that in the case of the front and back-wall cells, respectively, the metal counter electrode *3* and the blocking-layer *2*, or the contact metal layer *4* should be transparent enough to ensure a satisfactory functioning of the cells. As may be seen in fig. 1, the electrons move in the same direction as, and in the opposite direction to, that of the incident light in the front and back-wall cells, respectively.

---

[1]) Philips techn. Rev. 2, 13, 1937.

[2]) Philips techn. Rev, 4, 100, 1939.



a) b)

Fig. 1. Diagrammatic representation of the structure of a frontwall cell (*a*) and a backwall cell (*b*). *L* incident light, *1* semi-conductor, *2* blocking layer. The metal layer *3*, serving as a counter electrode is in *a* very thin and transparent, while in *b* it functions as base plate. In case *a*, however, the metal layer *4* making contact with the semi-conductor is the base plate and in case *b*, on the contrary, it is transparent.

At present two different main types of blocking-layer cells are being made, namely **cuprous oxide** and **selenium cells**; the latter are also made by Philips. The cuprous oxide cells are constructed either as front-wall cell or as back-well cell, selenium cells, on the other hand, only as front-wall cell. In order to make a **front-wall** cell of cuprous oxide (Cu_2O) one begins with a plate of cuprous oxide containing an excess of oxygen which lends it conductivity. This plate is then both *1* and *4* of fig. 1*a*. By reduction or ion bombardment the surface of this plate is treated in such a way that a blocking-layer *2* is formed on it of pure Cu_2O, which is an insulator. This latter is transparent and must finally be covered with a transparent metal layer *3*, towards which the electrons freed from the semi-conductor by the light will pass. The metal layer then becomes electrically negative and functions as counter electrode.

If on the other hand one wishes to make a **back-wall** cuprous oxide cell, one begins with a copper base plate whose upper surface is covered with a layer of cuprous oxide (Cu_2O) containing an excess of oxygen by being heated to a high temperature (1040 °C). This combination is then (acc. to fig. 1*b*) the counter electrode *3* and the semi-conductor *1* which are separated by a layer of pure Cu_2O which functions as blocking-layer *2*. Upon this is finally deposited a transparent layer of metal *4* which provides for the electrical contact with the semi-conductor *1*. The electrons freed by the light in the semi-conductor *1* now pass through the blocking-layer *2* to the metal base plate *3* and then return *via* the external circuit to the transparent metal layer *4*, which is in contact with the semi-conductor *1*.

As already mentioned, the **selenium cell** which is manufactured by Philips is a front-wall cell and the order of the layers is therefore as indicated in fig. 1*a*. Upon a base plate *4* of aluminium provided with a prepared surface a semi-conducting layer *1* of selenium is deposited. For this purpose the selenium has been prepared in the so-called grey modification by means of a heat treatment. The upper surface of this selenium is then provided with a blocking-layer *2* upon which a very thin transparent layer of metal *3* is finally deposited by cathode sputtering. This last layer functions as counter electrode. In most selenium cells the blocking-layer *2* is obtained from the semi-conductor *1* itself by causing a chemical reaction on its surface, which produces a non-conducting surface layer. Since, however, this blocking-layer does not completely cover the selenium, in the Philips selenium cell an additional **separate blocking-layer** is deposited, which layer has no chemical relation with the semi-conductor. As a result not only is a better blocking-layer obtained, but at the same time there is the advantage that the metal of the thin layer *3* which is deposited by evaporation on the blocking-layer *2* cannot react with the selenium, since the two layers are now entirely different chemically. The selenium cells with a separate blocking-layer can therefore be used much longer than those which do not have such a layer [3]).

Moreover, it is also an advantage of the separate blocking-layer that with it it is much easier to alter the characteristics ot the blocking-layer according to necessity. If, for example, it is made thicker, the photocell can be given a smaller capacity and at the same time the internal resistance can be increased, which may be desirable for certain applications.

Voltage and current of the blocking-layer photocell

When light passes through the transparent metal layer *3* and the blocking-layer *2* to the semi-conductor *1*, two different phenomena take place. Not only do electrons move from the semi-conductor through the blocking-layer to the metal counter electrode (**blocking-layer** photo-effect, as has already been explained), but at the same time the electrons freed in the selenium by the light produce a higher conductivity [4]) of this semi-conductor, which is called the **internal** photoeffect. Everyone suspects that there is a relation between these two effects, but it is still certainly a question whether or not the electrons which in the blocking-layer effect pass from the selenium to the blocking-layer are first freed in the selenium by the internal photoeffect. While the blocking-layer photoeffect occurs immediately upon illumination, the internal photoeffect shows inertia.

---

[3]) In the case of the blocking-layer **rectifiers** which are manufactured by Philips, **both** of these blocking-layers are therefore also present.

When the two electrodes (*3* and *4*) are connected externally, for instance by an ammeter, upon illumination the electrons pass from the counter electrode *3* through the measuring instrument, which usually possesses only a low resistance to the metal layer *4* which makes contact with the semi-conductor *1*. The direction of the electric current is there just the opposite, as is indeed indicated by arrows in fig. 1. In all types of blocking-layer cells, therefore, upon illumination the counter electrode *3* becomes negative with respect to the semi-conductor *1*. Due to this difference of potential over the blocking-layer *2*, the electrons will now have the tendency to return from the metal counter electrode to the semi-conductor. This is indeed the direction of good transmission for a rectifier. If there were no question at all of a returning current, one would be concerned only with a pure **photocurrent**, which would be directly proportional to the intensity of illumination on the photocell, as is represented by the straight line *a* in *fig. 2*. In addition to this, however, there is the oppositely directed **conduction current**, which may for example be represented by curve *b* in fig. 2 and which is found to vary approximately as the square of the voltage between the electrodes *3* and *4*. The total current through the external circuit thus finally takes on the shape represented by curve *c* in fig. 2.

Upon closer consideration, however, the concept just given is found to be still too much simplified. The conduction current *b* which flows in the opposite direction and is due to the voltage over the blocking-layer is not determined exclusively by this voltage, but is found to depend also upon the degree of illumination to which the photocell is exposed. The work necessary to take one electron from the metal counter electrode into the blocking-layer (the so-called **work function**) is somewhat smaller due to the illumination, so that for a given voltage the opposing conduction current *b* will be somewhat higher for the illuminated cell than for the non-illuminated cell.

In the foregoing we have continually spoken of a **current** which the photocell can send through an external circuit, but it may be of importance to know also what **voltage** it can deliver under different circumstances. If the voltage is measured with a voltmeter which consumes no current, or, more simply, if the voltage is calculated from the current and a sufficiently high external resistance, one obtains a voltage, when the illumination is not too intense, which is practically proportional to the intensity of illumination, because both the **photocurrent** and the counter **voltage** excited are then practically linear with that intensity. With more intense illumination, however, the counter voltage increases about quadratically with the illumination, so that the total voltage then increases less rapidly than proportional to the intensity of illumination.

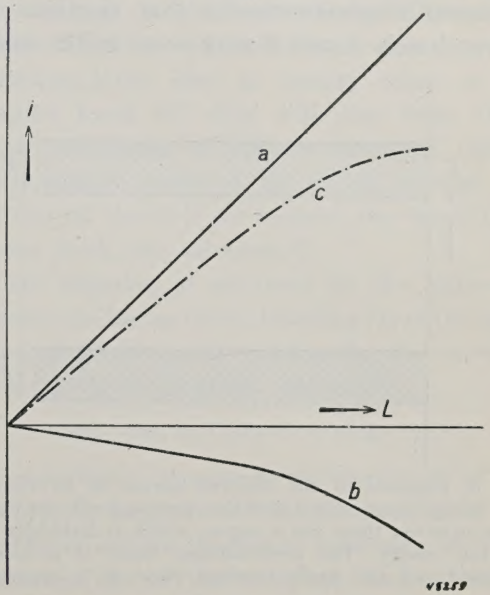


Fig. 2. The current generated *i* as a function of the illumination *L* of the photocell: *a* is the photocurrent which flows with no external resistance, *b* is the oppositely directed conduction current which is a result of the voltage over the blocking layer and *c* is the total current which flows through the external resistance.

Mechanism of the electron movement in blocking-layer photocells

We shall now attempt to explain somewhat more clearly the **mechanism** of the conduction by electrons in blocking-layer cells.

When free electrons are introduced from outside into a substance which is itself an insulator, the electrons can move in that substance, so that conduction is then possible in such an insulator. According to modern conceptions about the electrical structure of solid matter, however, it is not possible to introduce electrons with any arbitrary amount of energy from the outside. In a solid substance there are alternating regions of energy values (so-called energy bands) which are permitted or not permitted for the electrons which attempt to enter the substance. In *fig. 3* it is indicated diagrammatically that electrons of the energy bands A and B may occur in the solid, but the energy levels lying between them form a forbidden region.

---

[4]) Use is also made of this in technology. Thus for example a selenium cell, which in principle consists of a piece of selenium with two electrodes, can be put under voltage in a closed circuit. When this cell is illuminated the internal resistance decreases and the current consequently increases. (*Cf.* for instance Philips techn. Rev. 2, 13, 1937). By means of this change in current a relay may be closed. This type of selenium cell is often called a **resistance photocell**.

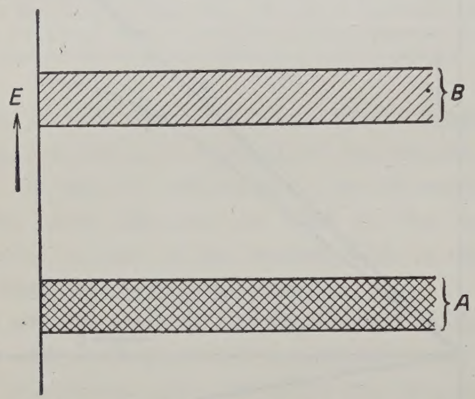


Fig. 3. Diagram of the electron states in an insulator. The energy bands A and B of the electron levels are permitted while between them lies a region which is forbidden for the electron energy. The cross-hatched band A is completely occupied and the single hatched band B is unoccupied.

In one definite energy state [5]), according to modern quantum theoretical conceptions, there may exist only one electron in every elementary cell of the crystal. An energy state may therefore be "occupied" or "unoccupied". If for example in the energy band A all the states are occupied, which has been indicated in fig. 3 by cross hatching, while the singly hatched band B is entirely unoccupied, it is not possible with the help of an electric field to give a small increase of energy to an electron, since all other states with energy values differing only slightly from the original value are already occupied and in order to attain an unoccupied energy state in band B a much too high amount of energy would be necessary! This is thus the case of an insulator in which no conduction of electricity can take place. However, by irradiation with light of a sufficiently high frequency ν an energy quantum $h\nu$ can, according to Einstein, be given to an electron existing in an energy state belonging to the fully occupied band A, by which that electron is brought into a higher state lying in the originally unoccupied energy band B. For electrons existing in energy states of the only partially occupied band B it is now possible, as a result of an external electric field, to pass over to other energy states which lie only slightly higher in the same band B. In this way the insulator has been made conductive by illumination, a phenomenon which was discussed in the foregoing as internal photoeffect.

With the help of the concept of energy bands here introduced we can now classify solid substances as to their electrical conductivity. If we are concerned with a full and an empty band, whose difference in energy values is large with respect to the average energy kT of thermal agitation, then the latter is not capable of causing the electrons to pass from the full to the empty band where they might be able to contribute to the conductivity. This is thus the picture, which we must imagine, of a perfect insulator. Now by adding a foreign substance to such an insulator it is possible also to create electron states whose energy values lie in the region C between A and B, as indicated in *fig. 4*. If the energy band C caused by the contamination now lies close below the unoccupied band B it is possible, that the average thermal agitation of the electrons will be capable of taking electrons from band C to states in band B, but this cannot take place on a large scale since there are only few energy states in band C available, due to the fact that the "contaminations" usually occur as a small percentage of the substance. In this case we are dealing with a semi-conductor.

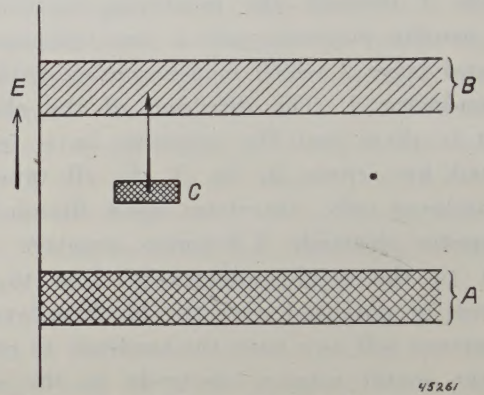


Fig. 4. The cross-hatched energy band A is occupied and the singly hatched band B is unoccupied. The likewise occupied energy band C is produced for instance by an impurity which makes the non-conducting substance a semi-conductor.

When, however, the occupied and the unoccupied bands pass from one to the other without interruption we are concerned actually with a partially occupied band B. In this case a continuous change in the electronic energy is possible and thus also electrical conduction. We have in this way arrived at a concept of a conducting metal. At the same time it is clear from this representation of the state of affairs that upon increase of temperature (larger kT) the average thermal agitation of the electrons will be better able to bring an electron from a lower occupied state into a higher unoccupied one, so that as a result of the partial occupation of the higher energy band B thus produced the electronic conductivity of insulators and semi-conductors will increase at higher temperatures, although it may in general be disregarded compared with the conductivity of metals.

---

[5]) Such a state is characterized not only by the value of the energy but also in general by various other quantities, such for example as the characteristic impulse moment of the electrons, the so-called "electron spin".

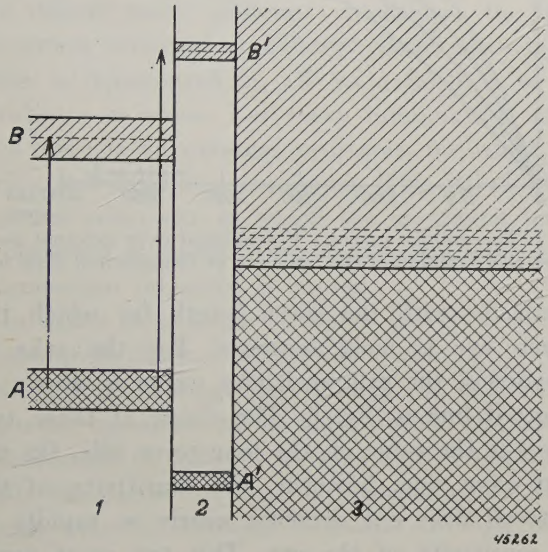


Fig. 5. Diagram of the position of the energy levels in the combination of a semi-conductor *1* with a metal *3* separated by a blocking layer *2*. A' and B' occupied and unoccupied energy bands, respectively, in the blocking layer. In the metal we are concerned with a continuum of permitted energy levels, which are completely occupied (cross-hatched) up to a certain limiting energy and above that only very sparingly occupied (single hatched).

For the combination of a semi-conductor and a metal which are separated from each other by a thin blocking-layer of an insulating material, we arrive, for the state of equilibrium, at approximately the circumstances shown diagrammatically in *fig. 5* for the position of the energy levels. In the case of a selenium cell, which is a front-wall cell as shown in fig. 1*a*, the light (incident from the right in fig. 5) penetrates through the thin layer of metal and blocking-layer into the semi-conductor and will there be able to take the electrons out of the occupied band A to the empty band B, so that those electrons will now increase the conductivity of the selenium (internal photoeffect). The electrons from the full band A of the semi-conducting selenium may, however, also be brought by the light into the still higher unoccupied band B' of the blocking-layer and in that way make that insulator slightly conducting. They then move through the blocking-layer to the transparent metal counter electrode and in that way produce the blocking-layer photoeffect. The amount of energy B'—A which an electron must receive from the incident light for this purpose is called the work function for the electron in the case of the blocking-layer cells and it is quite analogous to the work function of the electrons in the case of alkali-metal photocells, which in that case is necessary to take the electrons out of the alkali metal into the vacuum. The insulator in the blocking-layer photocell thus plays exactly the same part as the vacuum in the case of the alkali-metal photocell. When the photoelectrons once move in the blocking-layer into an energy state of the unoccupied band B', they will pass from there into the continuum of slightly occupied (singly hatched) energy states of the metal, so that the circuit can be closed if we connect the metal by a conductor with the selenium [6]).

For the emission of electrons by the light-sensitive semi-conductor to the blocking-layer the same equation of Einstein is valid as for the emission of electrons in alkali-metal photocells:

$$h\nu - h\nu_0 = {}^1/_2\, m\, v^2 = e\, V_m \quad . \; . \; . \quad (1)$$

In this equation ν represents the frequency of the light which is incident on the blocking-layer cell and ν_0 the minimum frequency necessary to bring the electrons out of the occupied band A of the semi-conductor into the unoccupied band B' of the blocking-layer. The remaining energy is thus possessed by the conduction electrons in the blocking-layer as kinetic energy: ${}^1/_2\, mv^2$. Finally V_m represents the minimum counter voltage which must be applied in order to prevent the electrons leaving the semi-conductor and passing to the blocking-layer; this is thus the so-called blocking voltage.

Since selenium absorbs visible light very strongly the latter can penetrate to only a small depth into selenium, so that the electrons are freed in the semi-conductor in a region which lies close to the blocking-layer. There is thus reason to suppose that the electrons which produce the blocking-layer photoeffect are freed from the semi-conductor by a photoeffect in the way just described, and that the mechanism described, at least in the main, satisfactorily accounts for the phenomena of conduction by electrons in blocking-layer photocells. It is, however, difficult to give a satisfactory explanation of the way in which these phenomena depend upon the temperature, and we shall not go into it here. We only call attention to the fact that, upon the application from the outside of a negative voltage on the selenium, a very strong photoeffect occurs in which phenomena of inertia appear, so that in this case there is obviously a close relation with the internal photoeffect which also shows a time lag.

---

[6]) If in this connection the question should be asked why electrons are taken by the light only out of the semi-conductor and not out of the metal into the unoccupied band B' of the blocking-layer, the answer is that in the metal itself such a large number of unoccupied states are present for the electrons which might take on such an energy from the light, that they will have practically no chance of just entering that band B of the blocking-layer, while on the other hand for an electron with that energy no permissible energy state is available in the semi-conductor, so that that electron must indeed be taken up in the unoccupied band B' of the blocking-layer.

Data of the selenium cell

In *fig. 6*, current characteristics are given of the selenium cells of 2 by 4 cm manufactured by Philips. It may be seen that, at least upon illumination with white light, with an external resistance of 10 ohms, there is a linear relation between the electric current and the intensity of illumination, which even at 2500 lux still shows no tendency toward saturation. With 100 ohms a slight saturation appears from 2000 lux onwards, due to the already discussed conduction current in the opposite direction. With still higher external resistance this saturation effect makes itself more and more clearly felt upon stronger illumination, so that with an external resistance of at least 1000 ohms one can indeed scarcely speak any longer of a linear initial section of the current characteristic. With an external resistance of for instance 11 000 ohms, even with an illumination of 2500 lux, the current becomes no higher than 15 μA, which fact can no longer clearly be shown in fig. 6.

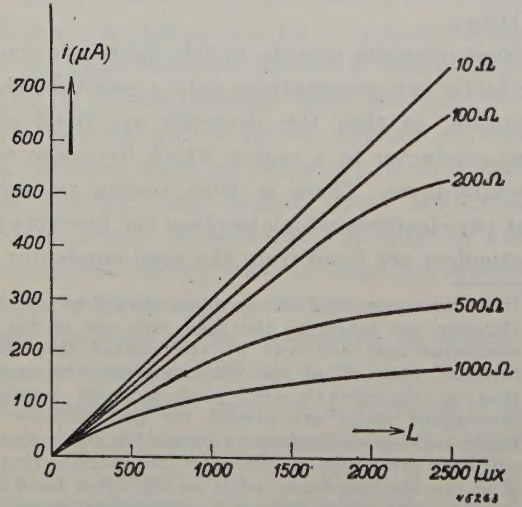


Fig. 6. Current i in μA as a function of the intensity of illumination L with white light at different values of the external resistance.

The behaviour of the voltage over external resistances of 500, 1000, 2000 and 11 000 ohms is shown separately in *fig. 7*.

In *fig. 8* the spectral sensitivity of the selenium cell is shown, *i.e.* the relation between the current which is produced upon illumination with a given energy of an arbitrary wave length and the current which is obtained with the same energy of illumination with a wave length of about 5500 Å. The latter is the wave length for which the selenium cell is most sensitive; this happens to coincide practically with the wave length for which the human eye is most sensitive. For the sake of comparison the eye-sensitivity curve is drawn as a dotted line in fig. 8. The shape of these two curves is the same on the long-wave side. On the short-wave side, however, the sensitivity of the photocell does not decrease nearly so rapidly as the sensitivity of the eye. This too great sensitivity of the cell for the shorter waves can easily be compensated by using suitable filters. In this way it is possible to imitate very well with the blocking layer photocell the spectral variation of the sensitivity of the eye. The selenium cell with filters

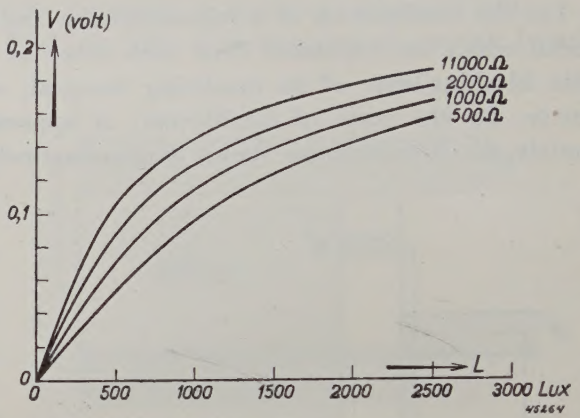


Fig. 7. The voltage in volts V generated over external resistances as a function of the intensity of illumination L in lux.

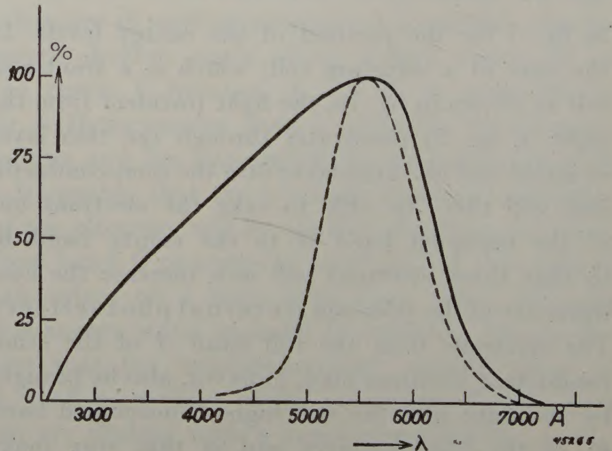


Fig. 8. The continuous curve represents the spectral sensitivity of the selenium cell in comparison with the curve of eye sensitivity shown as a dotted curve.

is therefore very suitable for use in the measurement of illumination intensities.

With a very low external resistance the current is found to be independent of the temperature, but at higher external resistances such a dependence does occur, since the resistance for the counter current depends upon the temperature. The size of the temperature coefficient of the current depends very closely upon the way in which the cell is constructed, but also upon the intensity of illumination. The current may vary by 1 to 10 per thousand per degree of change in temperature. The temperature coefficient is found to increase with the illumination intensity and with the external resistance.

In conclusion we shall briefly examine how great the power is which a blocking-layer photocell can deliver under different conditions. In *fig. 9* the power delivered at different illumination intensities is represented in μW as a function of the resistance in ohms. For every illumination intensity there is an external resistance for which the power delivered is a maximum. This value of the external resistance at which the maximum power is delivered is found to be smaller, the higher the illumination intensity is chosen.

In order to generate a higher power a number of photocells can be connected in parallel, so that with a given voltage one obtains the sum of the separate currents. As already stated, the photocells furnish in the first instance a photocurrent, while the photovoltage is a phenomenon which occurs over the external resistance through which the current flows. The cells must not therefore be connected in series, because in that case one only obtains a photocurrent equal to that of the least sensitive cell.

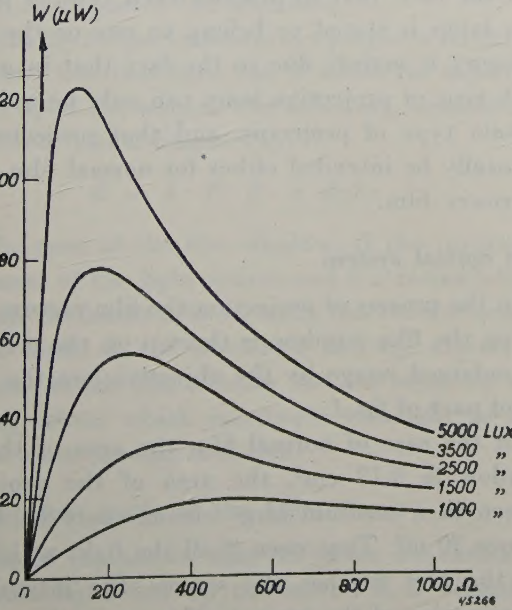


Fig. 9. The power in μW delivered by a Philips selenium cell as a function of the external resistance in ohms at different light intensities in lux.

If one converts radiant energy of the sun into electrical energy by means of a blocking-layer cell, the efficiency proves to amount to only about 2 percent. The power of a large cell with a surface of about 150 cm^2 in the full sunlight is found to be only just enough to make a small incandescent lamp burn or to drive a small motor; for the generation of higher electrical powers the blocking-layer photocell is not suitable.

Applications of the selenium cell

One of the most familiar applications of the blocking-layer cell is its use as photographic exposure meter. With a given photographic material the blackening attained by that material upon exposure during a certain time is proportional to the logarithm of the incident intensity. Now in the case of the blocking-layer cell, when an external resistance of about 500 ohms is used, the current generated is also practically proportional to the logarithm of the intensity of the incident illumination. We have already observed this in the character of the curves of fig. 6, while in *fig. 10* the curve for 500 ohms is plotted with a logarithmic scale of the illumination intensity and proves to be almost straight. We may therefore say that the electric current generated for a given illumination intensity is a suitable measure of the blackening of the photographic material, and that a good photographic exposure meter may be based upon this principle.

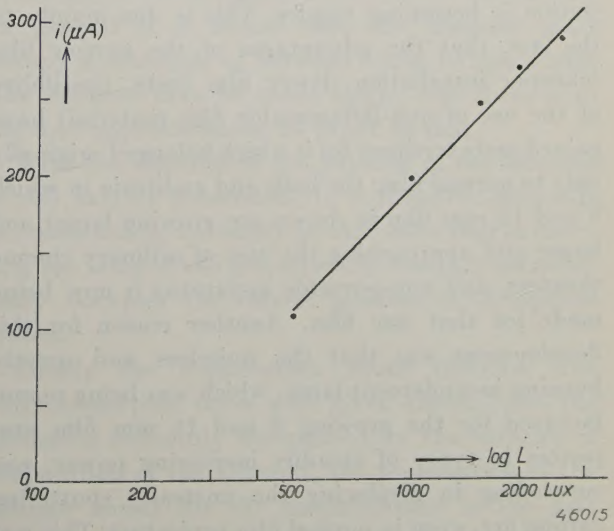


Fig. 10. The current with an external resistance of 500 ohms proves to be approximately a linear function of the logarithm of the illumination intensity in lux.

INCANDESCENT LAMPS FOR FILM PROJECTION

by Th. J. J. A. MANDERS.

621.326.73

Because of the characteristics of the optical system in which film projection lamps are used, the following requirements must be made of such lamps: The filament must have great brightness and must fill a given small area as completely as possible, the diameter of the bulb must be small and the lamp must be able to be adjusted automatically. On the basis of examples from among the numerous types of cinema and 8 and 16 mm film lamps manufactured by Philips, it is explained how these and other requirements are satisfied. Particularly the construction of the filament and the part played by the gas filling of the lamp are discussed from different points of view.

Cinema lamps and 16 mm film lamps

When in addition to the normal film 35 mm wide, used in cinemas, the narrower film (originally $9^1/_2$ mm, later 8 and 16 mm also) began to be used, it seemed at first as if not only the field of application but also the projection apparatus for the two types would differ in principle. The main problem in projection technique is that of passing the large light flux, which must be thrown on the screen, through the small film window, whose area in the case of 16 mm film is about one fifth, and in that of 8 mm film about one twentieth of that for normal film. Narrow film was therefore predestined for use in the home or in small auditoria with a small projection screen, where less light was needed. And while normal film was projected with the help of arc lamps and mirror condenser, the narrow film projector for the small audience worked exclusively with incandescent lamps and lens condensers. Only in that way could they be made sufficiently small and cheap, safe in use, easily transportable and fairly foolproof.

At the present time, however, the former sharp boundary between normal and narrow film projection is becoming vaguer. This is due mainly to the fact, that the advantages of the narrow film (cheaper installation, lower film costs, possibility of the use of non-inflammable film material) have gained more territory for it which belonged originally only to normal film: the halls and auditoria in which 8 and 16 mm film is shown are growing larger and larger and approaching the size of ordinary cinema theaters, and non-portable apparatus is now being made for that size film. Another reason for this development was that the noiseless and quietly burning incandescent lamp, which was being manufactured for the growing 8 and 16 mm film projectors in types of steadily increasing power, was succeeding in displacing the unsteady sputtering carbon arc, even in normal film projectors. This was possible to a certain extent in smaller cinema, but not in large theaters, where, because of its greater brightness, the arc lamp continued in use. During the last few years before the war the super high-pressure mercury lamp appeared as a new serious rival of the arc lamp [1]), and the intrusive narrow film has in the meantime also taken possession of this aid in order to win new territory. The field of projection in auditoria, which are not too large, remains in the undisputed possession of the incandescent lamp, either with normal or with 8 or 16 mm film.

In discussing incandescent lamps for film projection, we shall not, according to the above, be able to draw any sharp boundary between lamps for normal film — called cinema lamps — and lamps for narrower film. There is indeed a difference in lifetime: according to universally accepted standards the lifetime of cinema lamps must be 100 hours, while in the case of narrow film lamps 50, 25 or even only 10 hours is considered sufficient. This difference is, however, based more on tradition than upon logical grounds. In principle the requirements for cinema and for narrow film lamps are the same, and the fact, that in practice each type of projection lamp is stated to belong to one or the other category is mainly due to the fact that in general each type of projection lamp can only be used in a certain type of projector, and that projector will naturally be intended either for normal film or for narrower film.

The optical system

In the process of projection the film picture occupying the film window is thrown on the screen as an enlarged image by the objective; see the right-hand part of *fig. 1*.

In the case of normal film the area of the film window is 3.17 cm², the area of the projection screen in a medium large cinema may be for instance 20 m². Thus even if all the light which falls on the film reaches the screen, the intensity of illumination of the screen will be more than 60 000 times as small as that on the film window. If for example it is desired to have 50 lux on the screen (75 or 100 lux is usually demanded), the intensity of illumination on the film, taking into account the losses by reflection and absorption in the objective (about 50 percent), must amount to about 6 million lux.

---

[1]) See Philips techn. Rev. 4, 2, 1939.

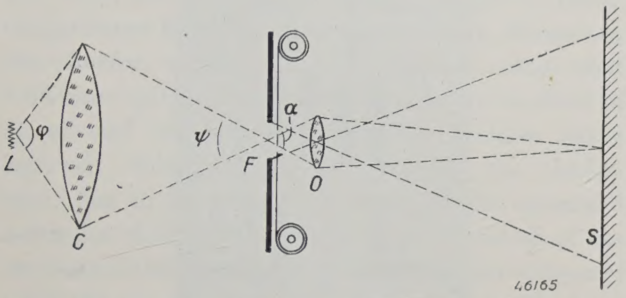


Fig. 1. The optical system used for film projection. The light from the filament of the lamp *L* is collected over a wide angle by the condenserlens *C* and focussed on or near the film window *F*. The film picture passing *F* is projected by the objective *O* on the screen *S*.

This enormously intense illumination is obtained by projecting the image of a very bright light source, in our case the filament of the projection lamp, on or near the film window by means of a condenser system; see the left-hand part of fig. 1. As may be seen in the figure the condenser has the function of collecting the light emitted by the source over a very wide angle φ and concentrating it on the film within a smaller angle ψ. The angle ψ should be equal to the angle α within which the objective can take up light from the film window. If $\psi < \alpha$, the objective is not entirely filled and if $\psi > \alpha$, unused light passes the objective.

Ordinarily the performance of a projector is characterized by the total light flux Φ which the objective throws upon the screen. The following formula can be derived [2]) in a simple way:

$$\Phi = k \cdot F \cdot B \cdot \pi \sin^2\alpha.$$

F is the area of the film window, *B* the (average) brightness of the light source and *k* a factor which takes all the losses into account, the chief of which are the losses in the condenser and the objective, the losses at the film window and the losses by the rotating sector which interrupts the light in the periods when the film is moved.

Requirements made of the light source

Formula (1) shows that in order to obtain a maximum light flux on the screen it is desirable to raise the brightness of the light source as high as possible. In addition the loss factor k must be kept as small as possible. In our case the losses at the film window constitute a large part of these losses. Due to the fact that the filament of an incandescent lamp is not a uniformly luminous plane, it may not be focussed exactly on the film window. If it were, the structure of that filament image would also be projected by the objective on the screen. The filament is therefore focussed farther away, in the objective or at a spot between objective and film window. In other words, the film window is not situated at the narrowest part of the beam of light produced by the condenser, and a larger or smaller part (70 percent for example) of this beam is cut off by the edges of the film window. The more uniform the brightness of the filament, the closer its image may be focussed to the film window and the smaller the loss at that point. The greatest possible uniformity of brightness of the filament is thus a second requirement.

Further requirements follow from a consideration of the condenser part of the projector. The light from the projection lamp is obviously used more economically the larger the angle φ compared to $\psi = \alpha$. If it is desired to limit the distance between source of light and film window to a certain length, in order to prevent the whole projector from becoming too large, a large angle φ means that the filament of the lamp must be able to be placed very close to the condenser, thus that the diameter of the envelope of the lamp must be made as small as possible. Furthermore with given values of φ and ψ a certain enlargement of the filament is obtained. Since the size of the image of the filament is fixed by the size of the objective to be filled, the minimum size of the filament is thus also prescribed. In general it is found that the filament must occupy a rectangle of between 5 and 10 mm width. The filament may be made larger, but there is in the first approximation no profit in that.

With the dimensions the position of the filament is also exactly prescribed. It is therefore desirable in constructing the lamp to take care that in inserting the lamp in the projector the filament automatically occupies the correct position in the optical system.

Finally the requirement about lifetime. The various accepted times were given above. Since the lifetime is ordinarily determined by the velocity of evaporation of the filament, the prescription of the lifetime means for the manufacturer in the first instance that he must choose a suitable working temperature. The brightness is also mainly given thereby. We shall see how the manufacturer, with the limited number of available degrees of freedom (nature and configuration of the incandescent body, gas filling), tries to increase the brightness as much as possible. In addition he must take care that the lamp also lives out its natural lifetime and does not succumb too soon to the mechanical vibrations to which it is exposed in the projector due to the jerking motion of the film transport. He must also take care that the light yield of the lamp does not decrease due to blackening of the envelope or deformations in the small, very rigidly fixed filament.

---

[2]) See the article referred to in footnote [1]).

Construction of the filament

The use of spiralized wire

For the construction of the filament, according to the above, there are three primary requirements: it must cover a certain rectangle of from 5 to 10 mm width, its brightness must be as uniform as possible and the average brightness as high as possible.

In connection with the first and second requirements it might be thought that a tungsten ribbon would be the ideal solution, such as is used in the ribbon lamps developed for photometry. A very uniform luminous surface could then be obtained in a simple way. This would, however, be at the expense of the third requirement: compared with the spiralized wire used in electric lamps for the same lifetime the working temperature and consequently the brightness of the ribbon must be chosen considerably lower. This is because of the fact that the ribbon has a comparatively larger surface for evaporation, while, moreover, the tolerances in thickness in the rolling of ribbon are larger than in the drawing of wire. The permissible temperature which must be calculated, so that the thinnest spots will not succumb before the desired lifetime is reached, is thus further depressed in the case of the ribbon.

Also in brightness obtained at a given temperature of incandescence the ribbon is inferior to the spiralized wire, due to the fact that with the latter a certain "black body effect" occurs [3]).

Thus for the sake of the third requirement, great brightness, a spiralized wire is used, and with it the prescribed area is filled as well as possible. For this purpose the spiral is usually divided into a number of sections which are assembled parallel to each other in a vertical plane, for instance as shown in *fig. 2*. The sections of the spiral will preferably be placed as close together as possible. In the first place the uniformity of brightness of the image of the filament is thereby improved or, speaking more accurately, its structure becomes finer, so that a slight lack of definition in the image is sufficient to render the structure invisible. The image of the filament can then be focussed at a spot closer to the film window. In the second place of course the average brightness of the image of the filament becomes larger according as the area covered by the non-luminous intermediate spaces between the wires is smaller.

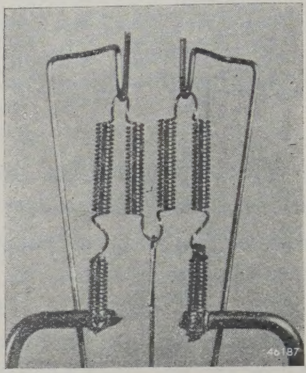


Fig. 2. Filament of a spiralized wire composed of four parallel sections of spiral.

Filling the filament area

Projection lamps, like most electric lamps for lighting purposes, are filled with gas which helps to prevent the evaporation of the filament and allows the filament to burn at a higher temperature and attain the same lifetime. Due to the gas filling, however, it is impossible to decrease the distance between the parts of the spiral indefinitely because of the necessity of avoiding breakdown between adjacent parts.

In spite of this limitation various means are employed to arrive as far as possible at a complete filling of the prescribed area.

In the first place it is obvious that the lamp should be made for a low voltage. The voltage between the parts of the filament and thus the necessary distance between those parts then automatically become small. In addition there are other, more important arguments for the choice of a low voltage. At a given power the current is inversely proportional to the voltage. A larger current corresponds to a shorter and thicker wire. A thicker wire may, for a given lifetime, burn at a higher temperature, since it offers a relatively smaller surface for evaporation of the tungsten. The lamps for low voltage therefore have a greater brightness for the same lifetime than those for higher voltage. In order to obtain an impression of the difference, let us compare a 110 V, 100 W lamp and a 30 V, 100 W lamp, both for a lifetime of 100 hours. The first has a filament 65 μ thick, the temperature of incandescence is 3000 °K, the brightness obtained 350 stilb; the second has a filament 155 μ thick, temperature 3100 °K, brightness 950 stilb. Moreover, the shorter, thicker wire has the advantage that a shorter spiral is sufficient, thus fewer sections of spiral and consequently less intermediate space between them, while in addition the thicker wire, due to its greater strength, can be spiralized around a thicker mandrel, so that the width of the sections of the spiral is more advantageous compared with the spaces between. It is even possible here, using a thick mandrel, to fill up the whole filament area very satisfactorily with a single spiral without subdivision.

---

[3]) Clean bare tungsten has an emission coefficient of about 0,5. Due to the fact that, in the case of the spiral, part of the radiation is reflected one or more times back and forth between the windings before it reaches the outside, the radiation takes on more the character of that from a black body, and the emission coefficient correspondingly approaches more closely the value of unity.

Apart from the heat losses through the leads, which are more important in the case of a short wire and high current (thick lead wires) and which prevent the voltage from ever being lower than about 15 V, a low voltage has only the disadvantage that for connection with the mains an intermediate apparatus is necessary, either a transformer, or, if the purchase price is more important than the efficiency of the projector, a resistance.

Since, because of this disadvantage, many manufacturers of projectors preferred to continue using high voltages, still other means of filling the filament area more uniformly are employed. A method which is used in all kinds of projection lamps (including those for low voltage) is the introduction of a spherical mirror behind the filament. This auxiliary mirror casts a real image of the filament in such a way that the images of the spiral sections fall exactly in the spaces between the actual spiral sections. The advantage in average brightness on the film window to be gained in this way may amount to 50 percent. The best uniformity will obviously be obtained when the intermediate spaces between the sections of spiral are not wider than the sections themselves, as is the case with the filament of *fig. 2*. If, however, the filament is one for a high voltage, the spiral must, on the one hand, be made thin in order for it to be strong enough with the thin weak wire which must be used, while on the other hand the spaces between the spiral sections must be large in order to avoid breakdown; compare such a filament for 220 V with one for 30 V (*fig. 3*). With such a rarefied filling of the area, as the figure shows, even the auxiliary mirror can only ensure a moderate uniformity.

A greater effect is obtained by doubly spiralizing the filament. This principle (sufficiently familiar from the incandescent lamps for ordinary lighting purpose ("Bi-Arlita")) improves the efficiency by decreasing the transfer of heat to the gas, permits the employment of a somewhat higher temperature for the same lifetime, due to a slight decrease in the evaporation, reinforces the already mentioned black-body effect and in our case also offers the possibility of filling the filament area about as well as if we were dealing with a low-voltage wire of the now spiralized single spiral. In this way it is possible, for example, to obtain the same average brightness with a 110 V-double spiral

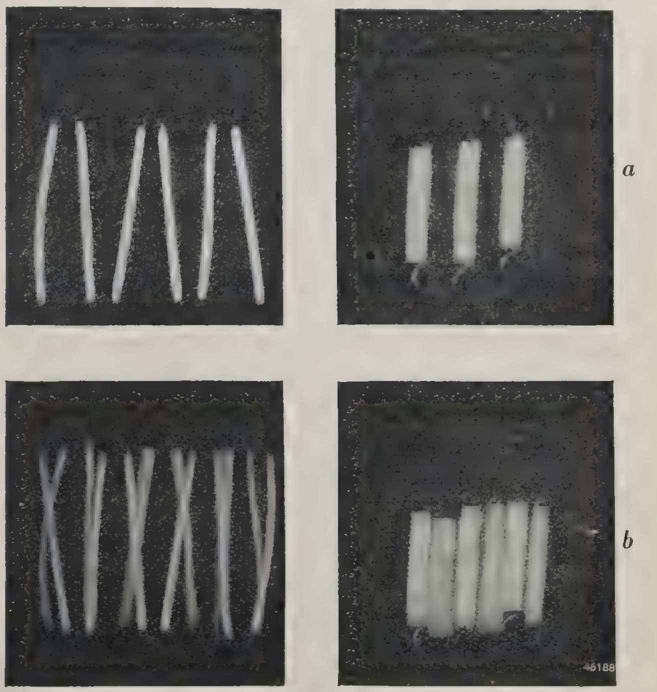


Fig. 3. Filaments of a 220 V narrow film lamp (left) and a 30 V narrow film lamp (right). *a*) The image of the filament without auxiliary mirror, *b*) with auxiliary mirror. In the case of the filament for the high voltage the filling of the area is unsatisfactory, even with the employment of the auxiliary mirror.

Table I.

Average brightness, without auxiliary mirror, of four different projection lamps for 100 W and a lifetime of 30 hours.

Voltage volts	Spiral	Dimensions of filament (mm²)	Average brightness (stilb)
30	single	5,2×4,5	1150
60	single	6,5×5	700
110	double	5,5×5	700
220	double	10 ×8	225

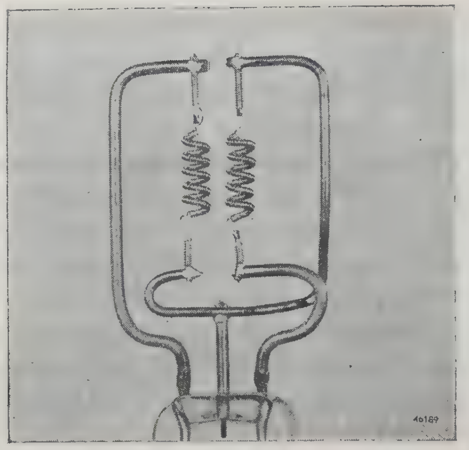


Fig. 4. Filament of doubly spiralized wire (110 V, 250 W). Note the connections of the two spiral sections: the current flows through both in the same direction, so that there is only half the lamp voltage between each corresponding pair of points in the two sections. Because of this the two sections may be situated closer together than when the two extremities at which the current is applied and taken off are situated side by side at the same height.

as with a 60 V single spiral; see *table I*. In *fig. 4* such a doubly spiralized filament is shown.

Finally the two-plane assembly is also much employed: the sections of spiral are not assembled in a single plane, but in two parallel planes in such a way that the spirals of one plane lie behind the intermediate spaces of the other, see *fig. 5*. It is clear that in this way very good uniformity and a very high average brightness can be obtained, especially when an auxiliary mirror is used in addition. As may be seen in *Table II*, these two-plane filaments are far superior to the others. In the case of one of the lamps here mentioned, by means of a combination of methods — two-plane assembly, low voltage and auxiliary mirror — a brightness of 4500 stilb has been obtained. This is certainly the highest value that has yet been attained with incandescent lamps. For the sake of comparison it may be mentioned that the spiral of an ordinary 100 Dlm 220 V lamp for the home has a brightness of 680 stilb, a fluorescent tubular lamp 0.3, the carbon arc 20 000, the high-intensity arc 80 000, the water-cooled super high-pressure mercury lamp (SP 1000) in the axis of the discharge 45 000 stilb.

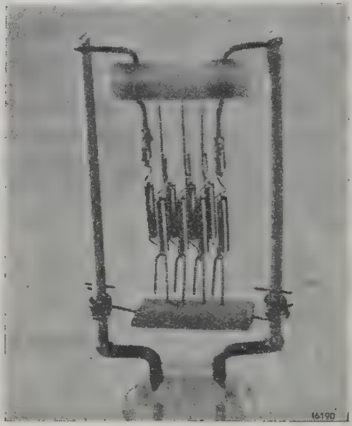


Fig. 5. Two-plane filament, seen obliquely from above. The system of hooks which protect the spiral sections on the upper side from lateral deviations is fastened rigidly to the leads, the corresponding hooks on the lower side, on the contrary, can move up and down in a vertical direction thanks to an arrangement of two tubes which slide over the leads (so-called sag arrangement, see below). The insulators should be noted between the hooks and the leads: these are made of a special ceramic material prepared and shaped in the factory and then fired with the hooks inserted.

Table II.

Data of several Philips cinema and narrow film lamps

Power watt	Voltage volts	Lifetime*) (without auxiliary mirror) hours	Form of filament mm^2	Dimensions of filament	Total light fluxlumen	Average brightness *) (with auxiliary mirror) stilb
100	110	100	one plane	9 × 6,2	1 700	480
100	30	50	one plane	5,2 × 4,5	2 300	1650
250	110	100	one plane	10 × 8	5 950	1200
250	50	50	one plane (double sp.)	7 × 6	6 000	2300
500	110	100	one plane	12 × 10,5	12 700	1650
500	110	25	two plane	7,8 × 7,6	12 500	3500
750	15	100	one plane	8,5 × 9,5	20 000	4100
750	110	25	two plane	9,5 × 9	19 800	3800
900	30	100	one plane	11 × 12,5	25 000	3000
900	30	100	two plane	11,5 × 8	25 000	4500
1200	110	10	two plane	12 × 12	36 500	4200

*) The life time without, the average brightness with auxiliary mirror.

The measures described all have the effect of promoting a greater concentration of the filament and a higher average brightness. Their effect is, however, partly cancelled by a more irregular distribution of temperature, of which they themselves are the cause. Since the temperature of the hottest part of the filament determines the lifetime, the more irregular the distribution of the temperature the lower the average temperature must be chosen for a given lifetime. Because of this, part of the gain in brightness must be sacrificed. This effect — the unfavourable ratio between highest and average temperature — is already appreciable upon passing over from single to double spiral; it is stronger in the case of the transition from single-plane to two-plane filament; and it is further accentuated when an auxiliary mirror is added because of the fact that part of the reflected radiation falls upon the filament again and raises its temperature locally. If the decreased lifetime is accepted, it would actually have been possible to make some gain in brightness without the mirror by raising the temperature of the filament. Independent of this controversy, however, remains the gain obtained in the projector due to the fact that with more uniform brightness the image of the filament can be focussed closer to the film window.

In speaking of the non-uniformity of the brightness of the filament we have had in mind until now the more or less periodic variations which are encountered as one passes across the spiral sections and the intermediate spaces of the filament or even along the length of a spiral over the windings. Superposed on this is a gradual decrease in brightness from the middle to the edges of the filament, caused by the fact that the edges cool more rapidly because of the unhindered lateral radiation, see *fig. 6*. Of itself such a not too rapid decrease toward the edges, which is reproduced in the illumination of the film picture and consequently also in the illumination on the screen, is not a serious difficulty. It necessitates, however, a certain correction in the statement made above about the required size of the filament: it is indeed an advantage to make the filament somewhat larger than the minimum size. When this is done only the middle, hottest part of the filament is used, whose average brightness may be considerably higher than the rest. And in practice this advantage is still further increased by the fact that at a given voltage a larger filament corresponds to a higher power and consequently a higher current, a thicker wire and thus a higher permissible temperature for the same lifetime.

Finally a few words about the gasfilling. The voltage at which, for a given distance between sections of the spiral, breakdown may happen depends upon the gas. For normal lamps a mixture of nitrogen and argon is used. The argon causes less loss of energy by the conduction of heat and therefore improves the light yield per watt, but it decreases the voltage for breakdown. As for film projection the specific light yield is less important than a maximum concentration, these lamps are filled with pure nitrogen, which allows of the sections of the spiral being placed closer together without risk of breakdown. It must, however, be kept in mind that even without that danger a certain distance would have to be maintained between the spiral sections in order to avoid any chance of short circuit between the sections due to possible deformations from mechanical vibrations, heat expansion or recrystallization. It should be mentioned here, that it has only become possible due to the progress in the manufacture of tungsten during the last ten years to diminish the deformations due to recrystallization upon ageing of the lamp to such an extent that it became possible to construct two-plane filaments. In the case of single-plane filaments also, it is of course desirable that the spiral sections shall remain straight during the life of the lamp. But the requirement is not so forcible in this case, since only upon deformation of the spirals in one direction is there chance of short circuit, while with two-plane filaments deformation in almost every direction may result in short circuit.

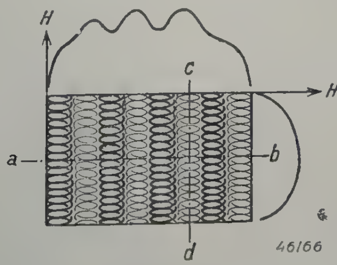


Fig. 6. Distribution of brightness in the filament image obtained with a two-plane filament with auxiliary mirror. Above, the distribution measured along the line *a-b*, to the right, that along *c-d*. The brightness is measured with a photocell of about half the width of the spiral image. It is plotted in arbitrary units. (See: G. Mili, J. Motion Pict. Eng. **28**, 168, 1937).

Position of the filament

In order to satisfy the above-mentioned requirement that upon inserting the lamp in the projector the filament will automatically be situated at the correct spot, each lamp is provided with an inner socket and an outer socket, the latter of which can only be placed in the lamp holder of the projector in a single definite way (so-called centering socket). In a model of the given projector the lamp with inner socket is first placed in exactly the correct position with respect to the outer socket, which is checked by projection of the incandescent filament on a small screen with indicating marks, see *fig. 7*. The inner socket is then soldered into the outer one.

It is essential that the adjustment described should be carried out while the lamp is burning normally, because in the cold state the filament is in general in a different position than in the hot state, due to the fact that the leads and hooks and the filament itself expand when hot. A similar method is used for the locating of the auxiliary mirror on the wall of the envelope. The adjustment of the filament with respect to the mirror takes place during the fusing in of the pinch, simply by observing the mirror image of the filament. In the cold state filament and mirror image must be shifted mutually a certain distance in order to fit together satisfactorily when the filament is hot (see *fig. 10*).

It is of course necessary that all parts of the filament remain sufficiently firmly in place during the life of the lamp and are not deformed or warped by possible tensions in the material. How such tensions may occur is shown in *fig. 8*. The spiral in two sections here shown as an example must be secured by a hook at the lower side to prevent lateral deviation. Due to the fact that the thin filament becomes hot and cold much more rapidly than the thick leads, upon switching on the spiral has a slight play for several seconds, upon switching off on the other hand the spiral (and the hook) would be under a high tension for several moments. In order to prevent this in the case of the more complicated filaments (two-plane systems) a so-called sag arrangement is employed; *cf. fig. 5*. The hooks are not rigidly fixed, but are able to move slightly up and down vertically, while still giving the spiral sections sufficient support against lateral deviations.

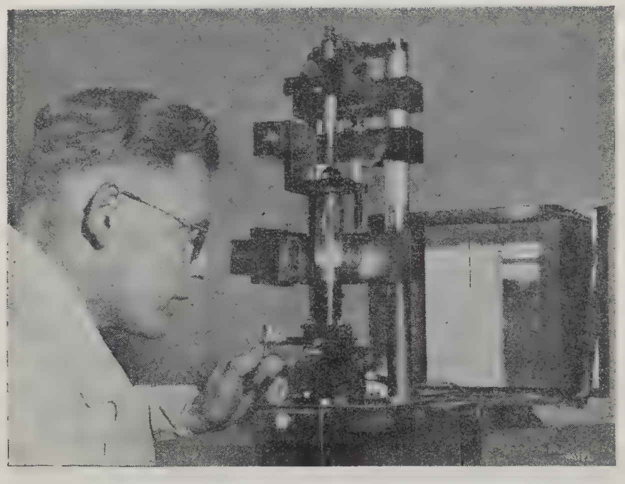


Fig. 7. The instrument for adjusting the filament with respect to the centering socket. To the right on the screen may be seen the filament images, one of which is projected directly on the screen and the other, seen perpendicular to it, is projected *via* a prism. By turning and shifting the lamp the images must be made to coincide with a drawing on the screen. The correct position of this drawing with respect to the lamp holder is checked from time to time with a normal lamp.

Thanks to these measures and to the previously mentioned absence of deformation in the spirals due to recrystallization, all parts of the filament retain their correct positions during the whole lifetime, and no decrease in light occurs in the projector due to gradual loss of adjustment.

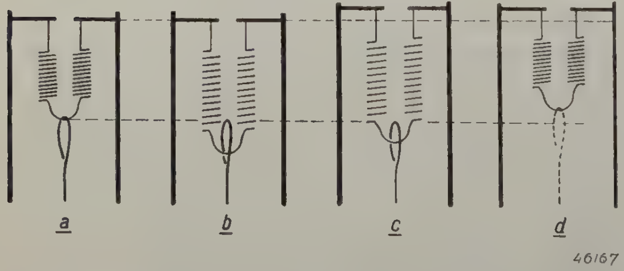


Fig. 8. The expansion and contraction of a two-section filament and the leads upon switching on and off. *a*) Cold state; *b*) immediately after switching on: spiral hot, leads still cold; *c*) in operation; *d*) directly after switching off: spiral cold, leads still hot.

The envelope

The requirement mentioned at the beginning, that it must be possible to place the filament as close as possible to the condenser, leads to the fact that the envelope, usually of a tubular form, must be given as small a diameter as possible. A limit is set by the fact that the wall of the envelope would become too hot if it were too close to the incandescent filament. This can also be enhanced by the blackening which normally appears after the course of time on the wall due to the evaporated tungsten, which is deposited on the wall and which absorbs part of the radiation and thus makes the envelope still hotter. Three methods are available of arriving at a small diameter of the envelope. In the first place the envelope is made of a special hard glass which has a very high softening point, namely 750 °K. In the second place, especially in projectors for 8 and 16 mm film, a forced cooling is employed by means of a fan. In the third place attempts are made to prevent the deposition of the evaporated tungsten on the wall in the neighbourhood of the filament. This is indeed desirable in any case because the blackening is accompanied by a gradually increasing reduction of the light during the lifetime of the lamp.

The gas with which the lamp is filled plays an important part in these phenomena. While on the one hand it has the disadvantage that it becomes warm itself by heat conduction from the filament and therefore heats the envelope more strongly, on the other hand it has a favourable effect as far as the blackening is concerned, because of the fact that currents occur in the heated gas which carry the evaporated tungsten along, so that most of it is not deposited close to the filament but at some other spot. This effect can be reinforced by introducing nickel gauzes or plates above the filament in the rising gas current, upon which the tungsten particles are deposited (see *fig. 9*). In this way it has been possible to limit the diameter of the envelope of lamps for 8 and 16 mm film for 1000 W and 1200 W to the same as that for a 750 W lamp, namely only 38 mm. Still more effective is the employment of the so-called hanging construction; *fig. 10*. The lead wires, in this case above the filament, which become very hot close to the filament, function as a sort of chimney, which causes

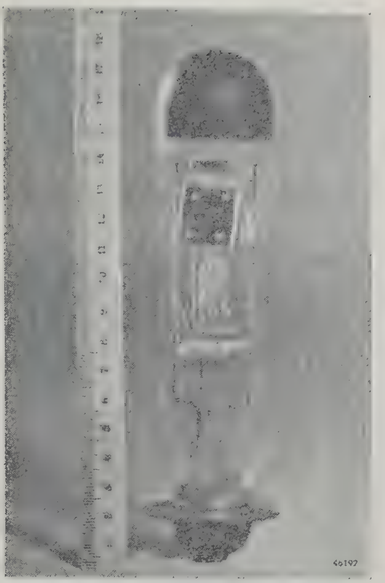


Fig. 9. Philips 16 mm film lamp, 110 V, 1000 W. Above the filament are two nickel plates along which the rising gas flows and upon which the particles of evaporated tungsten have an opportunity to be deposited. The upper part of the envelope is sprayed black in order to prevent light escaping through the ventilation holes of the projector into the auditorium.

the rising current of gas to pass first along the axis of the lamp. In the spherical enlargement at the top of the envelope the gas takes on a whirling motion in which practically all of the tungsten carried with it has the opportunity of being deposited on the wall at that spot. The descending, cooled gas, in which there are only few tungsten particles left, deviates from the wall again slightly above the filament and flows toward the filament (*fig. 11*). Therefore at the height of the filament the wall remains almost entirely free of blackening, while also there is no heating of the wall here by passing hot gases. In that way in the case of the 750 W lamp shown in fig. 10, which had to operate without artificial cooling, the diameter could be reduced to almost the same value (44 mm) as in the case of the above mentioned lamps with forced cooling.

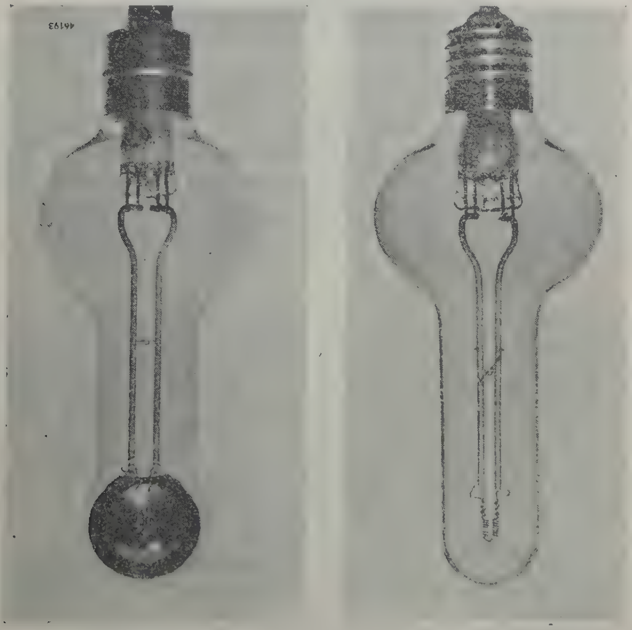


Fig. 10. Two Philips cinema lamps 15 V, 750 W with the so-called hanging construction. The left-hand lamp has an auxiliary mirror of evaporated aluminium on the inside of the envelope.

The gas currents described can be made visible in a very interesting way by constructing such a lamp with a tube fused on, through which, while the lamp is burning, a small amount of air can be admitted to the lamp. The tungsten particles coming from the filament are then immediately oxidized to a white powder whose motion can be followed with the eye. The well-known method of Toepler can also be followed, in which the current phenomena are made visible by means of the accompanying very slight local variations in the index of refraction. These methods are very useful in the investigation of the shape of envelope which is most suitable for obtaining the desired course of the currents of gas.

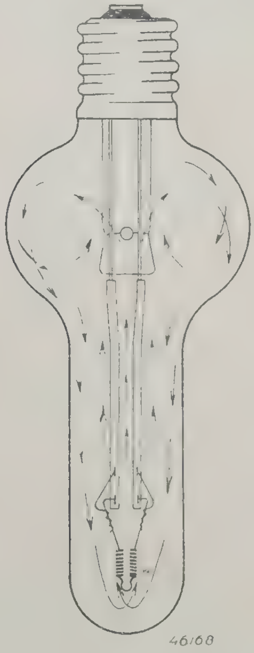


Fig. 11. Gas currents in the lamps of fig. 10.

In still another way the gas helps in protecting the envelope. If, due to an incipient blackening or possible insufficient cooling, the envelope should become soft at a certain spot, then at that spot it would be slightly compressed by the external pressure if there was a vacuum or diminished pressure inside the lamp. It would thus come nearer the filament and thus become still hotter and still softer, and this accelerated process would quickly lead to the destruction of the lamp. Now by filling the lamp with a gas at such a pressure that when hot there is a slight excess pressure in the lamp, the compression of the envelope is made impossible. The excess pressure can be employed without any danger because of the small size and thus the strength of the envelope. The relatively high pressure gives the additional advantage of a still greater increase in the permissible working temperature and an increase in the breakdown voltage in the gas.

Finally it may be mentioned that for some 8 or 16 mm film projectors, with a very compact construction, lamps have been made in which the filament is brought especially close to the wall by excentric placing in the envelope. In the case of the 400 W lamp shown in *fig. 12*, for example, the distance between filament and wall is only 7 mm. This is made possible by the employment of the hanging construction and by an arrangement for cooling in the projector housing in such a way that the part of the envelope lying opposite the filament is the most strongly cooled. For the rest it is striking that in this lamp the spiral sections are horizontal. The vertical position is usually chosen in order to prevent short circuit between the sections in case of any slight sagging of the spiral during use. The horizontal position has the advantage that the light emitted in the axes of the envelope can easily be used for the scanning of the sound track of the film, so that a separate lamp is not needed for that purpose. In the case of the lamp in fig. 12 use is made of this possibility, which exists indeed only with the hanging construction, since in the ordinary, standing construction the light along the axis direction would have to pass through the upper parts of the envelope, which are most subject to blackening.

Although in projectors for normal film the dimensions of the apparatus are much less critical than in the case of narrower film, here also for other reasons the diameter of the envelope must be kept as small as possible. Since in general these projectors work with a mirror condenser, the auxiliary mirror is placed in front of the lamp and thus cuts off a part of the beam of light which the condenser mirror throws on the film. The auxiliary mirror should therefore be small, and since the mirror is placed on or outside the envelope, the envelope itself must be small.

Attempts have indeed been made to avoid this necessity by placing a very small mirror inside the envelope very close to the filament. This mirror had to be made of molybdenum because of the intense heat. The relatively low coefficient of reflection of molybdenum (60 percent in the cold state, about 75 percent at white heat, compared with 95 percent for a mirror of aluminium deposited on the wall) and the practically unavoidable loss of adjustment upon becoming hot, however, make the advantage of this method more or less illusory, especially since the hanging construction has made it possible to reduce the diameter of the envelope very much at the position of the auxiliary mirror (in the lamp of fig. 10, left, it amounts to only 60 mm.)

The lamps in operation

We have already pointed out that the lamp must be resistant to the mechanical vibrations to which it is exposed when the projector is in use. These vibrations are composed of many components of different frequencies and are different for each type of projector. It is therefore impossible to indicate a lamp construction which could be used in all cases. Attempts are rather made empirically during the development of the lamps to render a lamp insensitive to the special vibrations of the type of projector for which it is intended. For this purpose the lifetime is determined of test models of the lamp and a normally working model of the projector, and the construction of the leads and hooks are altered until the desired lifetime in operation is fully achieved. This work is made more difficult by the requirement that the leads and hooks must not stand in the way of the light beams which the filament emits towards the condenser and auxiliary mirror.

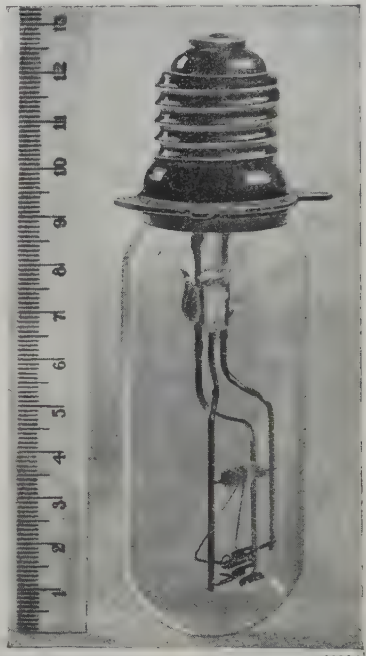


Fig. 12. Philips 16 mm film lamp 30 V, 400 W, with hanging construction and with excentrically placed filament, whereby the distance between filament and wall of the envelope is reduced to 7 mm. The spiral sections are horizontal in this case, the light emitted downward is used for scanning the sound track of the film. The sections of the spiral do not lie in a flat plane but in a cylindrically bent plane, whereby a similar effect is obtained as in the two-plane assembly.

The nominal lifetime can only be attained if the lamp is burned precisely at the prescribed voltage or current. If the lamp is calculated for a certain voltage, an increase in voltage of 1 percent decreases the lifetime by 14 percent, while on the other hand a decrease in voltage of 1 percent causes a decrease in brightness of 4 percent. With a lamp for a certain current the corresponding variations upon change in current are still larger, namely 25 percent in lifetime and 8 percent in brightness. The difference between the two kinds of lamps consists in the fact that, in choosing the dimensions of the former, account is taken of a gradually decreasing current strength during the life of the lamp, due to the fact that the filament is becoming thinner by evaporation, and a consequently increasing resistance; in the case of the second lamp a gradually increasing voltage is discounted in the dimensions for the same reason. Since as a result of this, the brightness will decrease slightly in the first lamp and increase in the second lamp, the use of lamps for constant current is always considered desirable for projectors in order to compensate a decrease in light due to blackening of the walls and deformation of the filament. Since, however, according to the above in the Philips lamps this decrease in light yield has been reduced to practically zero, the burning of the lamp at constant voltage would be more recommendable in connection with the abovementioned slighter sensitivity to variations in the operating condition, and also because the "gain" in light upon the ageing of the lamps for constant current is of course only obtained at the expense of a lower temperature of incandescence and brightness in the new lamp. In any case the voltage or current stamped on each lamp should be exactly maintained during its use if full profit is to be had with respect to brightness and lifetime. Most projectors are provided with a measuring instrument (although it is unfortunately not always reliable) for controlling voltage or current.

A few words may be said about the efficiency of the lamps. As can be seen from *table II*, it is very high, thanks to the high temperature of incandescence. The lamps with a lifetime of 25 hours have as a rule a specific light yield of 25 lm/W, the 1200 W lamp in the table with a 10 hour lifetime even has more than 30 lm/W, which begins to approach the efficiency obtained with gas-discharge lamps. If one considers the unusually uneconomical way in which the light is used in the projector, the high efficiency mentioned appears to be a waste of effort. A balance of the light flux excited in the lamp for different projectors shows that in the end a maximum of only 1 to 2 percent of it reaches the screen as effective light. Nevertheless, the high light yield of the lamp is of very great practical advantage, not so much because of the lower current costs — these are negligible compared with the cost of the film — but because of the smaller amount of heat developed. This byproduct must be removed from the projector without the occurrence of undesired high temperatures at any spot, and in particular the heat contribution from the radiation which must be concentrated on the film window may not cause at that spot too great overheating. This heating is indeed not so dangerous in the case of film of non-inflammable material, and it is less intense in the case of the lens condenser used for narrow film than with a mirror condenser, since due to the dispersion in the glass the lens has the fortunate property of focussing the infra red rays to an image of the filament considerably farther away. On the other hand, however, in the case of just these narrow film apparatus every limitation of the heat developed is welcome in order to be able to reach a reasonable light production and still retain a small and convenient construction.

THE MEASUREMENT OF REVERBERATION

by W. TAK.

534.844.1

In this article several theoretical considerations about reverberation in rooms and auditoria are discussed. The phenomenon of reverberation is considered as the sum of a number of damped characteristic vibrations. Due to interference, fluctuations are superposed upon the mainly exponential variation. The variation in intensity depends upon the shape and dimensions of the room, upon the distribution and the properties of the absorption material, upon the position of the observer, upon the position of the source of sound and upon the frequency spectrum of the sound produced, but when these factors remain constant it is completely reproducible.

Several principles are indicated according to which apparatus can be constructed which makes it possible to study the variation in intensity during the reverberation and to characterize it by one or more parameters. A more detailed description of the apparatus will be given in a subsequent article.

It is known that in a room where sound is produced the intensity does not immediately decrease to zero after the source of sound has ceased to act. Depending on the circumstances, a certain time is necessary before the sound dies out. This phenomenon of reverberation has great influence, not only on the intensity of the sound, but also on the intelligibility of speech and the quality of music, and is therefore important in judging the acoustics of the room. It is not only a question of the length of time during which the reverberation lasts, but just as much of the variation of the intensity as a function of time, *i.e.* of the character of the reverberation. This character is not, however, a property of the room in question only, but it is also determined by the nature and the position of the source of sound and the position of the observer. The influence of the duration of the reverberation on the intelligibility of speech and the analysis of music have been discussed previously[1]). We shall now go more deeply into the causes which determine the character of reverberation and indicate several principles by which it can be studied experimentally.

Characteristic vibrations of a closed space

The study of the vibrations of the air in closed spaces has shown that such a space possesses a series of characteristic frequencies each of which corresponds to a definite mode of vibration of that space.

We shall begin with the consideration of a simple case. We imagine a source of sound situated between two opposite plane walls of a rectangular room. The source emits a plane sound wave toward the front and rear with a wave front parallel to the walls mentioned above. A system of "stationary" waves will then occur between these walls, which is most strongly developed when the source emits a single frequency such that the distance between the walls is a multiple of a half wave length. These frequencies, which are characterized by the formulae

$$a = p\cdot\frac{\lambda}{2},\quad \lambda = \frac{2}{p/a},\quad \nu = \frac{c}{\lambda} = \frac{c}{2}\cdot\frac{p}{a},\qquad (1)$$
$$p = 1, 2, 3 \ldots,$$

where a is the distance between the walls in question, λ the wave length, ν the frequency and c the velocity of sound, form a series of "characteristic frequencies".

In *fig. 1* the modes of vibration of the air (amplitude of the air velocity and the sound pressure as a function of the position) are represented for p = 1, 2, 3, 4. The walls are here assumed to be hard, so that they reflect the sound almost completely. In fig. 1 the spectrum belonging to this series of characteristic vibrations is also given. Actually

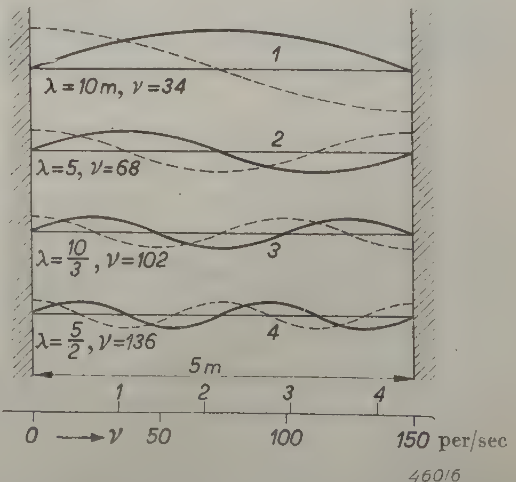


Fig. 1. The first four characteristic vibrations of a column of air between two parallel walls. Beneath the figure the spectrum of the corresponding characteristic frequencies is given.

---

[1]) Philips techn. Rev. 3, 65, 1938.
Philips techn. Rev. 3, 143, 1938.
Philips techn. Rev. 3, 368, 1938.

the frequency spectrum in a rectangular room is more complicated; in the first place because there are three pairs of opposite walls, each with its own distance apart, and further because plane waves also occur which are not propagated parallel to one of the walls.

The general formula which must then be used instead of (1) is the following:

$$\nu = \frac{c}{2}\sqrt{\frac{p^2}{a^2}+\frac{q^2}{b^2}+\frac{r^2}{h^2}}, \quad \lambda = \frac{2}{\sqrt{\frac{p^2}{a^2}+\frac{q^2}{b^2}+\frac{r^2}{h^2}}}, \qquad (2)$$

$$p, q, r = 1, 2, 3, 4. \ldots,$$

where a, b and h represent the length, width and height of the room and p, q and r are arbitrary numbers.

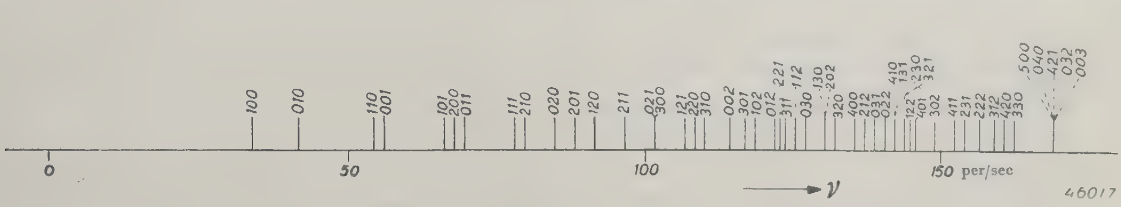


Fig. 2. Part of the frequency spectrum of a rectangular room with the dimensions $a = 5\text{m}$, $b = 4\text{m}$, $h = 3\text{m}$.

In *fig. 2* the frequency spectrum is given, as an example, for a room for which $a = 5\text{m}$, $b = 4\text{m}$ and $h = 3\text{m}$. It may be seen from this figure that the number of vibrations in a given frequency interval increases rapidly with increasing frequency. This can also be proved as follows. If in a three-dimensional system of coordinates one constructs the points for which

$$x = \frac{cp}{2a}, \quad y = \frac{cq}{2b}, \quad z = \frac{cr}{2h},$$

those points form a regular lattice in space with the average density:

$$\frac{8abh}{c^3}.$$

Furthermore:

$$\nu = \sqrt{x^2+y^2+z^2},$$

so that the distance from the point to the origin represents the frequency. The number of frequencies in the interval ν, $\nu + \Delta\nu$ is equal to the number of points lying in an octant of a spherical shell included between the spheres with radii ν and $\nu + \Delta\nu$, whose volume amounts to

$$\tfrac{1}{2}\pi\nu^2\Delta\nu.$$

With a sufficiently large value of ν this number can simply be set equal to the volume of the spherical shell, multiplied by the average density of the frequency points. The following is then found for the number of characteristic vibrations in the interval ν, $\nu + \Delta\nu$:

$$\tfrac{1}{2}\pi\nu^2\Delta\nu\cdot\frac{8abh}{c^3} = 4\pi\frac{abh}{c^3}\cdot\nu^2\Delta\nu.$$

For the above-mentioned room, in the neighbourhood of $\nu = 1000$ and when $\Delta\nu = 1$ c/s, the number of characteristic vibrations is already about 25.

What happens when the source of sound is now switched off? While previously at every point in the room there was a vibration with the frequency (or frequencies) of the source, we now observe a motion which can be conceived as a superposition of a number of characteristic vibrations, whereby especially those characteristic vibrations contribute to the sound image whose frequency differs little from the frequencies present in the source. Due to the absorption always present, these vibrations will be damped, so that the reverberation may therefore be conceived as *the sum of a number of damped characteristic vibrations.*

Intensity of the reverberation as a function of the time

Let us assume, by way of hypothesis, that in the closed space considered, only one of the characteristic vibrations, with the frequency ν_e, is active. At every point in the space pressure and velocity are sinusoidal functions of time. The sound energy, consisting of a potential part which is proportional to the square of the pressure, and a kinetic part proportional to the square of the velocity, is a periodic function of the time, which also depends upon the position of the observer. If there is no damping, the energy per unit volume ($\bar{E}$), averaged over the space, is constant. If there is slight damping the total energy $E \cdot V$ decreases with the time according to the formula: ²)

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\bar{E}\,V) = -\frac{c\,\bar{E}\,A}{4},$$

---

²) For the deduction of this formula see Philips techn. Rev. 3, 65, 1938.

so that:

$$\overline{E} = \overline{E}_0\, e^{-kt}, \text{ where } k = \frac{cA}{4V}.$$

V is here the total volume, E_0 the average energy per unit of volume at the moment $t = 0$, c the velocity of sound, $A = \int a\, \mathrm{d}O$, where $\mathrm{d}O$ represents an element of the wall and a the absorption at that spot. The sound energy thus decreases according to an exponential law. In the presence of damping the sound *pressure* at every point in the room will be an exponential function of the time, so that:

$$p = p_0\, e^{-k't} \sin 2\pi\nu_e t, \text{ where } k' = \frac{k}{2} = \frac{cA}{8V}.$$

This is represented in *fig. 3a*. The absolute value of the sound pressure may also be considered and averaged over a small number of periods. The average sound pressure $\overline{p}$ thus obtained also varies according to an exponential function of the time:

$$\overline{p} = p_0\, e^{-k't}.$$

In fig. 3*b* and *c*, $\overline{p}$ and $\ln \overline{p}$ are represented as functions of t.

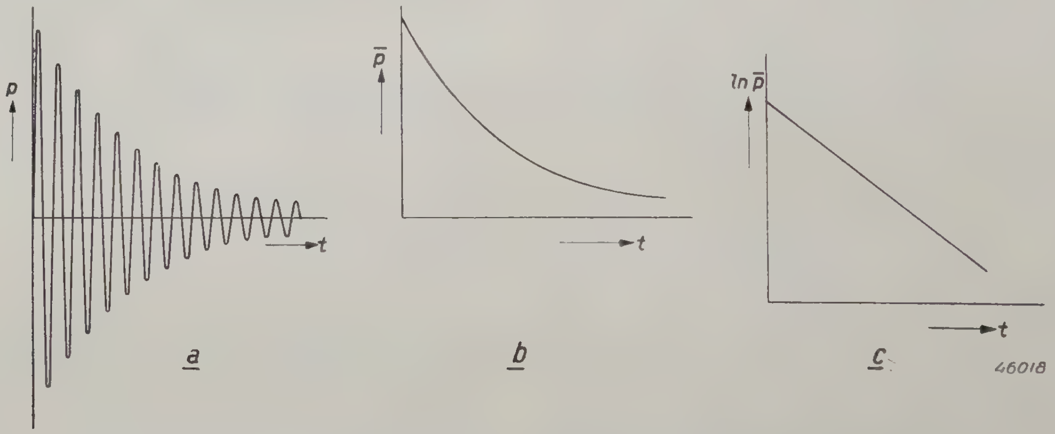


Fig. 3. Plotted as a function of t: *a*) the sound pressure p for a single damped vibration, *b*) the average sound pressure $\overline{p}$, *c*) $\ln \overline{p}$.

If there is more than one characteristic vibration active, each characteristic vibration individually will die out exponentially. Since, however, the frequencies differ, these vibrations will cause beats by interference, so that the pressures will exhibit fluctuations. When a single tone is emitted in a room by means of a loudspeaker, upon interruption of the sound the amplitudes of the excited characteristic vibrations will be smaller the larger the differences between the frequencies, and thus only those characteristic frequencies will be noticeably active whose frequencies differ only slightly from the given frequency. In a room in which the absorbent materials are distributed uniformly over the surface, the damping constants of these characteristic vibrations can in general be set equal to each other, because while the absorption coefficient depends upon the frequency, in a narrow frequency region it varies only slightly. The interference picture will now be quite simple and the amplitude modulation which occurs as a result of the interference will be clearly observable (*fig. 4a* and *b*). If the tone emitted by the loudspeaker is replaced by a more composite sound, the interference picture will exhibit a different character because a larger number of characteristic vibrations are then excited whose beats partially compensate each other, whereby the fluctuations, while more numerous, are less pronounced (*fig. 5a* and *b*). The damping constants can now vary more widely; this will in general have little effect on the variation in the intensity as long as the sound emitted does not cover a very extensive part of the spectrum (for instance several octaves). If the latter is, however, the case, and the wall absorbs for example

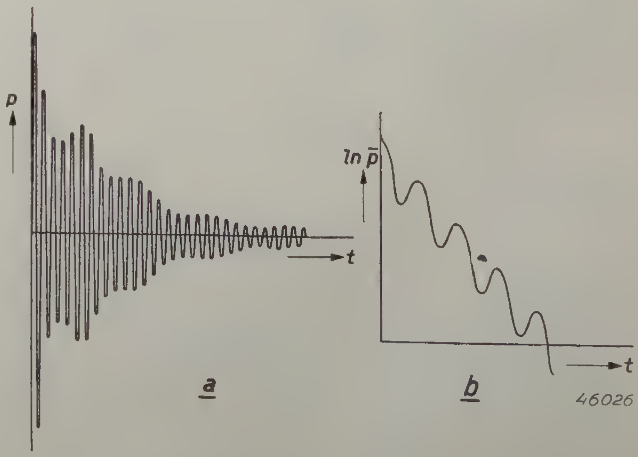


Fig. 4. Plotted as a function of t: *a*) the sound pressure p upon superposition of two damped characteristic vibrations with different frequencies, but with the same damping, *b*) $\ln \overline{p}$.

only the low frequencies, then due to the rapid dying out of the low excited frequencies the room will have a shrill sound. If on the other hand absorption in the high frequencies is predominant, the room has a hollow sound.

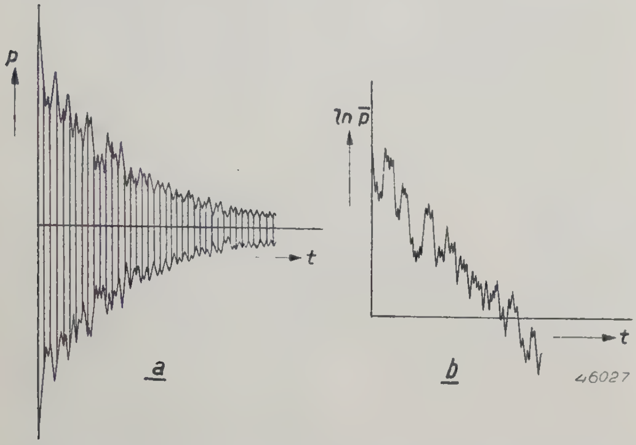


Fig. 5. Plotted as a function of t: a) the sound pressure p upon superposition of a large number of characteristic vibrations, b) ln $\bar{p}$.

The situation is different when the absorbent materials are distributed very irregularly over the surface, for instance in the case of a room with hard walls, whose floor is covered with a soft carpet. Here the vibrations whose waves planes are parallel to the walls will be much less damped than those with horizontal wave surfaces. In such a room characteristic frequencies with very different damping will occur at the same time. In the variation of the pressure with the time, apart from the fluctuations, a superposition of different exponential functions can be recognized, because the strongly damped characteristic vibrations first die out, and only later the weakly damped ones. A similar situation is encountered when a strongly absorbent room is connected by an opening with a room with slight damping, for example a "soft" room which is connected by an open door with a "hard" corridor. In that case it is advisable to consider the characteristic vibrations of each room separately. The characteristic frequencies will be slightly affected by the opening, which acts as a "coupling". The characteristic vibrations will only be altered in as much as those of the hard room penetrate slightly into the soft room, and *vice versa*. In this case also in the combined room, characteristic vibrations with practically the same frequency but with very different damping will exist at one and the same time. If the source of sound is situated in the room, the corridor will resonate faintly. An observer in the room hears chiefly the strongly damped characteristic frequencies of the room, followed by a faint, more lasting reverberation of the corridor (*fig. 6a*). An observer in the hard room, however, will notice chiefly the slowly dying sound in that room, while the quickly damped sound of the soft room, which initially reached his ear, will quickly escape his attention (fig. 6*b*).

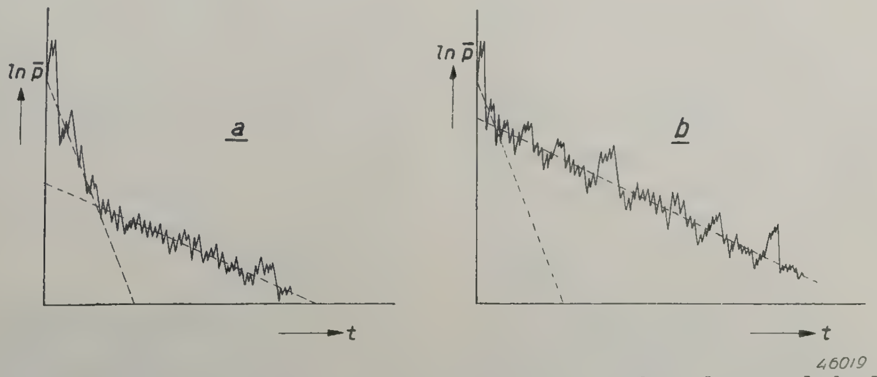


Fig. 6. Plotted as a function of t: a) ln $\bar{p}$ in a space composed of a damped room and a hard corridor for a point of observation in the room, b) the same for a point of observation in the corridor.

Reverberation time

We have seen that in the simplest case the average pressure varies as an exponential function of the time:

$$\bar{p} = \bar{p}_0\, e^{-k't}.$$

This variation can now obviously be characterized in a simple manner. If we now define the reverberation time as the time t_{60} during which the intensity decreases by a factor 10^6 (60 db) (and thus the pressure by a factor 10^3), we find for this time t_{60}:

$$10^{-3} = e^{-k't_{60}},$$

$$t_{60} = 2.3\,\frac{3}{k'} = \frac{13.8}{k}.$$

or, since $k = cA/4V$,

$$t_{60} = 0.16 \frac{V}{A},$$

where V is expressed in m^3 and A in m^2 of open window. In the case of an *exponential* drop the t_{60} measured is a parameter which completely determines the character of the reverberation. We shall therefore call it the reverberation time of the exponential reverberation.

If one has a diagram in which the variation of $ln\ \overline{p}$ (or $2\ ln\ \overline{p}$) with the time is given, the reverberation time can be determined therein by a simple construction. In *fig.* 7 this construction has been performed.

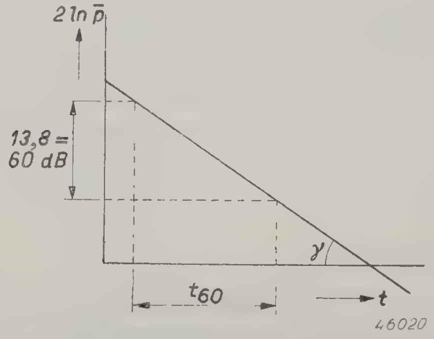


Fig. 7. Determination of the reverberation time in the case of an exponential drop of the intensity. The situation of the right line has been rendered by the measurements. It follows from the graph that tan t_{60} = 13.8 cot γ.

In more complicated cases, for example when strong fluctuations are superposed on the exponential variation, there is of course no simple relation between the reverberation time and the character of the curve. In that case it is obvious that an average exponential curve can be drawn, disregarding the fluctuatons, and a value of t_{60} for that curve can be found. We shall call this quantity the average reverberation time. It is true that the phenomenon is not thereby completely characterized, because the fluctuations have been disregarded, but it gives, nevertheless, the general character of the reverberation.

If the reverberation consists of the superposition of different exponential functions, each with its own constant k', perhaps combined with fluctuations, the obvious method is first to replace the curve by an average curve without fluctuations and to analyze that curve into its purely exponential components. If one then determines t_{60} for the separate components, one obtains a number of reverberation times which together to some extent fix the general character of the reverberation.

Quantitative investigation of reverberation

There are various experimental methods of investigating reverberation. The oldest is that of Sabine, by which only the reverberation time is determined. For that purpose the time was measured which elapsed between the moment at which the source of sound ceased to function (this might only take place when a stationary sound intensity had been reached) and the moment at which the sound became inaudible. This method was later improved by the use of an entirely automatically working apparatus which determines electrically the time in which the intensity of the signal decreases by 60 db [3]).

We have seen, however, that the reverberation depends very much upon different factors and that it may have a very capricious variation, so that the sole determination of the reverberation time cannot in this way give a picture of the true variation of intensity. At the present time use is made of recording meters, which have the advantage that the exact behaviour of the reverberation is recorded, making it possible to draw an average curve and to determine from it one or more average reverberation times, while, moreover, from the curve obtained conclusions may be drawn about the acoustic properties of the room being studied [4]). The principle of these recording meters is simple. After being amplified the voltage of the microphone is connected to the poles of a potentiometer, which is then automatically regulated so that the voltage taken off remains constant. The mechanical regulator at the same time drives a style which records the position of the potentiometer contact on a roll of paper. The necessary high velocity of recording, in connection with the unavoidable mechanical inertia, makes the whole mechanical part of this apparatus extremely delicate and difficult to transport. We have therefore developed two methods which are not subject to this objection and we shall discuss them in the following.

The first method we have called the method of "exponential amplification", the second that of the "exponential time base".

Method of exponential amplification

For the sake of simplicity we shall consider the case where the sound pressure varies purely exponentially, so that:

$$p = p_0\, e^{-k't} \sin 2\pi\nu t.$$

When we observe this quantity by means of a microphone, an amplifier and an electron-ray oscillograph, and arrange the apparatus in such a way that after the interruption of the source of sound the amplification increases according to an exponential function of the time e^{+at}, then, when a is chosen equal to k', the influence of the damping will just be cancelled by that of the amplification. The result is a simple sine curve with a constant amplitude, while for $a > k'$ and $a < k'$ the amplitude respectively increases and decreases with the time (*fig. 8a, b* and *c*). It is clear that we can also apply this to those cases where fluctuations are superposed on the general exponential behaviour.

---

[3]) M. J. O. Strutt. De Ingenieur 4 E, 20, 1932.

[4]) *C.f.* Philips techn. Rev. **3**, 65, 1938.

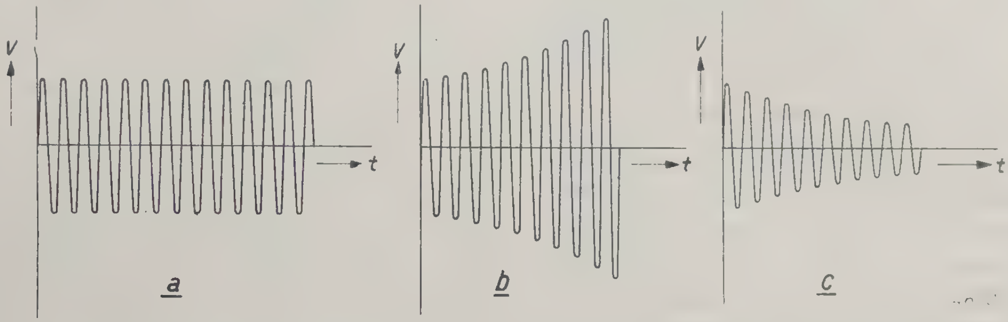


Fig. 8. The output voltage V of the exponential amplifier is plotted as a function of t for the case of a single damped characteristic vibration, when a) $a = k'$, b) $a > k'$, c) $a < k'$.

The fluctuations are, to be sure, not then eliminated from the result, but with a suitable choice of a the average amplitude can be kept constant (*fig. 9*). The whole can be made visible by connecting the output voltage of the "exponential amplifier" to the vertical diffraction plates of an electron-ray oscillograph. By synchronisation of the time base with the apparatus which regulates the switching on and off of the source of sound, it is possible repeatedly to project the image of the reverberation on the screen. Since the variation of the intensity in a given case, with the same arrangement of microphone and source of sound, is entirely reproducible, a constant image appears on the screen. By varying a, that arrangement can be found with which the image on the average has a constant amplitude. In that case $a = k'$, and from that the average reverberation time can be deduced, while at the same time the character of the reverberation can be studied in the image observed. In some cases (for example the above-mentioned combination of soft room + hard corridor) the reverberation consists of several terms with different values of k'. From the oscillogram it may then be seen that at several values of a the amplitude can be made constant for a part of the image (*fig. 10a* and *b*), from which the different reverberation times present simultaneously can be deduced. Here also the method described shows clearly the general character of the reverberation together with the details.

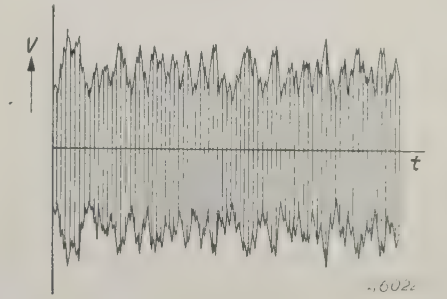


Fig. 9. The output voltage V for the case of a large number of damped characteristic vibrations.

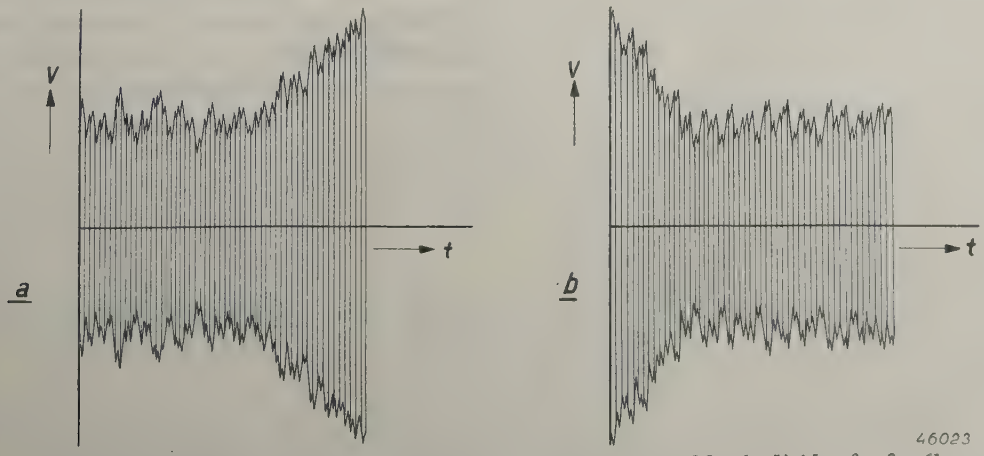


Fig. 10. The output voltage V as a function of t, a) for the case of fig. 6a, b) *idem* for fig. 6b.

When the intensity of the reverberation decreases, there comes a moment at which the deviation on the screen becomes just as large as the deviations which are the result of unavoidable extra noises and of the noise of the amplifier. The intensity will then no longer decrease and at the end the image on the screen will always show an increase in the amplitude. In order to avoid false conclusions the apparatus is so adjusted that upon reaching the noise level the curve is interrupted and the signal switched on again.

The method of the exponential time base

In this method the AC voltage from the microphone is connected *via* a rectifier with a smoothing arrangement to the plates for vertical diffraction of an electron-ray oscillograph. To the plates for horizontal diffraction a voltage is applied which does not increase linearly with time, as is customary (linear time base), but which varies according to an exponential function $e^{-\beta t}$. The movement of the light spot on the screen is then given by the functions

$$y = a\, e^{-k' t},$$
$$x = b\, e^{-\beta t}.$$

If we eliminate t from these, we obtain:

$$\left(\frac{y}{a}\right)^{\beta} = \left(\frac{x}{b}\right)^{k'},$$

$$y = C\, x^n, \qquad C = a\, b^{-n}, \qquad n = k'/\beta.$$

The curves represented by this equation (*fig. 11*) are in general sorts of parabolas. There is, however, one of them which has degenerated into a straight line ($n = 1$), which is the case when $k' = \beta$. By the adjustment of β, therefore, an image can be obtained in which the fluctuations take place about a straight line. Depending on the character of the reverberation, this adjustment is more or less sharply defined. In any case it is also possible to study the character of the reverberation precisely. One objection to this method is that only a relatively small part of the whole phenomenon of reverberation can be observed at one time. After a decrease in intensity of about 20 db the voltage applied to the vertical plates becomes so small that the change in it can no longer be clearly observed, due to the size of the light spot. This difficulty can be met by increasing the amplification, so that for $t = 0$, y falls outside the screen. Only at the moment $t = t_1$ at which the light spot, after having fallen vertically, again becomes visible on the screeen, is the horizontal time base switched on (*fig. 12*); the curve appearing on the screen then relates to the part of the reverberation for which $t > t_1$. The moment t_1 is determined by the amplification, so that in this way the behaviour of the reverberation can be studied during different parts of the reverberation time, especially also the case where the reverberation consists of the sum of a number of exponential functions. It is our intention to discuss these apparatus in a subsequent article and to give details of their construction and of the results obtained with them.

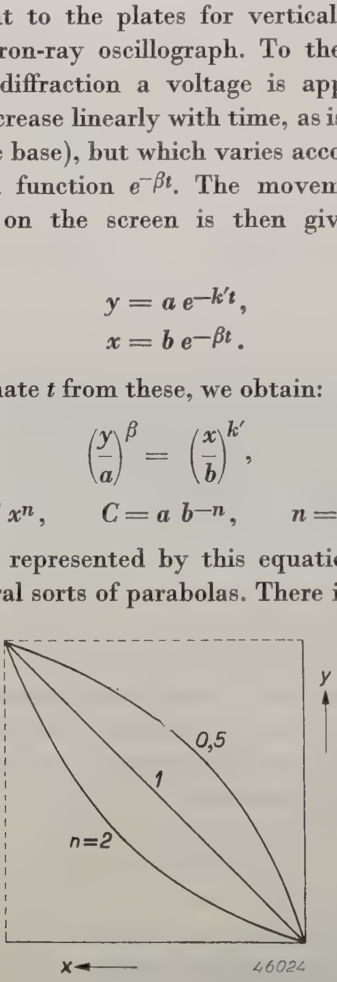


Fig. 11. The curves $y = Cx^n$ with $n = k'/\beta$ as parameter.

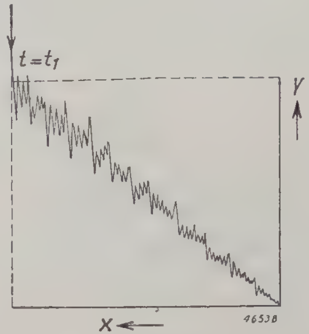


Fig. 12. The image of that part of the reverberation for which $t > t_1$, obtained by the method of the exponential time base.

COMPARISON OF FREQUENCY MODULATION AND AMPLITUDE MODULATION

by Th. J. WEIJERS.

621.396.619

As a continuation of a previous article on frequency modulation [1]) the circumstances are here discussed under which frequency modulation offers advantages over amplitude modulation for radio broadcasting. It is possible as a consequence of this to extent the frequency spectrum to be transmitted to the highest audible frequencies and there is no need in reproducing music to strain the dynamics. An amplitude limiter must be introduced in the receiver, while, moreover, it is desirable to use a push-pull detector. Further, the frequency sweep must be chosen much larger than the low-frequencies to be transmitted. In order to obtain good quality of reception also the deformation in transmitter and receiver should be kept extremely small. Only on short waves with wave length of not more than a few meters can a sufficiently large frequency sweep be used, and therefore only for such wave lengths is frequency modulation preferable to amplitude modulation.

In a previous number of this periodical [1]) it was explained what is meant by frequency modulation and amplitude modulation, and how in principle with these methods the modulation takes place in the transmitter and the detection in the receiver. We shall now compare the two methods of modulation, and in particular note their relative advantages and disadvantages.

Although amplitude and frequency modulation both make it possible to obtain an undistorted reception of high quality, provided transmitter and receiver, including microphone and loudspeaker, satisfy high requirements, they behave quite differently with respect to interferences: with frequency modulation under certain conditions a much lower noise level can be obtained. In this article we shall explain under which conditions the advantage can be obtained and how it appears in the quality of the reproduction. According to the manner in which the noises appear, we must distinguish two cases. During the modulation of the desired signal the interferences are to a large extent masked by the desired sound; during a pause in the modulation the interfering sounds are in general most annoying. We shall therefore in the first place ascertain what audible interferences occur when in a frequency-modulation receiver, as the simplest case, a sinusoidal interference is added to the unmodulated sinusoidal carrier. We shall then discuss the results of such an interference when the carrier is modulated. Finally we shall deal with the influence on the reception of the interferences due to noise and atmospherics, electrical apparatus and other transmitters.

The case in which a sinusoidal interference of small amplitude and the desired modulation are present alternately.

We shall now ascertain the influence of a sinusoidal interference added to the carrier at a moment when the latter is not modulated. For the sake of simplicity we shall confine ourselves for the present to the case where the amplitude of the interference is small compared with the amplitude of the carrier, for instance smaller than 1/10.

In *fig. 1* the unmodulated carrier *a* and the sinusoidal interference *b*, as well as their sum *c*, are plotted as a function of the time. From this figure it is clear that the amplitude of the resulting oscillation *c* is not constant, but that it varies with a frequency equal to the difference between the frequencies of the two oscillations *a* and *b*. The positions where the curves pass the zero axis are also changed, so that the distances between them are no longer equal. Thus, due to the interference, the carrier in question is given an amplitude modulation as well as a frequency modulation. The amplitude modulation has a depth equal to

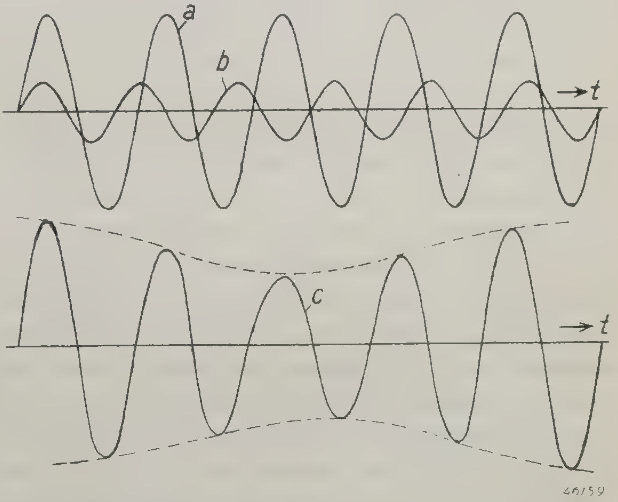


Fig. 1*a*) Unmodulated oscillation.
b) Interfering oscillation with small amplitude and a frequency somewhat higher than that of the oscillation *a*. For the sake of clearness the interference *b* is not drawn < 1/10, but about half as large as the main oscillation *a*.
c) The sum of *a* and *b*; from the form of *c* it can immediately be seen that the disturbance of *a* by *b* amounts not only to an amplitude modulation but also to a frequency modulation.

---

[1]) Philips techn. Rev. 8, 42, 1946.

the ratio of the amplitude of the interference b to that of the unmodulated carrier a.

In order to determine the sweep of the frequency modulation the various oscillations are plotted in a vector diagram in *fig. 2*. The unmodulated carrier a, whose amplitude we assume to be equal to 1 and which therefore can be represented by $\sin \omega_0 t$, is represented by the stationary vector OP. The interference b which can be represented by $\sin\,(\omega_0 + q_s)\,t$, where $s \ll 1$ and q_s is the difference between the angular frequencies of carrier and interference, is indicated by the vector PQ. This latter vector has the length s and rotates about the point P with an angular velocity q_s. The end point of this vector thus describes a circle with the radius s and centre P. The resultant oscillation c is now represented by the vector OQ moving back and forth, whose starting point O remains fixed, but whose end point Q describes the previously mentioned circle. From this it again appears that the amplitude of the resultant oscillation varies between $1-s$ and $1+s$, thus that the interference causes an amplitude modulation with a depth of modulation s. The instantaneous angular frequency of the resultant oscillation is equal to the instantaneous angular velocity of the vector OQ with respect to the time line OA which rotates about O with an angular frequency ω_0. In order to find this angular frequency we first determine the angle φ between OP and OQ. For that purpose we drop the perpendicular QR from Q on OP. The angle α between OP and PQ is $q_s t$, so that $QR = s \sin q_s t$. Since $s \ll 1$, OQ does not differ much from 1, and thus $\sin \varphi$ is approximately equal to $QR/OQ = s \sin q_s t$. Since in our case the angle φ always remains small, we may set $\sin \varphi = \varphi$, so that $\varphi = s \sin q_s t$. The angular velocity of OQ is $\mathrm{d}\varphi/\mathrm{d}t = s q_s \cos q_s t$; with respect to the time line OA, which rotates with an angular velocity ω_0, the instantaneous frequency of the resulting oscillation c is: $\omega_m = \omega_0 + s q_s \cos q_s t$. The frequency sweep caused by this interference is therefore $s q_s$, the product of the ratio s between the amplitudes of the interference and the carrier and the frequency difference q_s between the two oscillations.

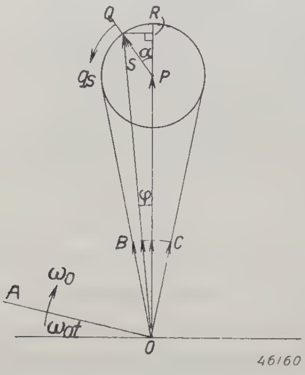


Fig. 2. Vector representation of the oscillations of fig. 1. The stationary vector OP with the length 1 is the unmodulated carrier a with a frequency ω_0 at which the time line OA rotates clockwise; PQ with a length s is the interfering oscillation b with a frequency $\omega_0 + q_s$, so that this vector rotates counter-clockwise around the point P with the frequency difference q_s. The resultant OQ is the modulated oscillation c. The amplitude limiter provides that the vector for the limited oscillation moves back and forth between the positions OB and OC.

We shall now examine how the receiver reacts to an oscillation of the form just discussed. The human ear can only receive sound impressions in a limited frequency region; the auditory limit lies at about 15 000 cycles per second. The angular frequency [2]) corresponding to this auditory limit, which thus amounts to $2 \cdot \pi \cdot 15\,000$, is called q_a.

For the sake of simplicity we shall assume that the frequency spectrum is cut off sharply at this frequency q_a by the low-frequency part of the receiver, the loudspeaker and the ear, so that higher frequencies are not heard and lower frequencies are perfectly heard.

In an amplitude-modulation receiver the detector provides a low-frequency oscillation with a frequency equal to the difference q_s between the frequencies of the two oscillations a and b and an amplitude proportional to the amplitude s of the interfering oscillation b. After amplification this is made audible by the loudspeaker. The frequency modulation which the carrier has obtained from the small sinusoidal interference has no influence on this process of detection; only the amplitude modulation caused by the interference makes itself felt in the same way as a desired amplitude modulation. If the carrier a is sinusoidal amplitude-modulated with a depth of modulation m, the low frequency detected desired signal is proportional to m. Upon maximum modulation $m = 1$. The ratio of the amplitude of the interference during a pause in the desired modulation to the maximum amplitude of the desired signal is therefore at the loudspeaker the same as at the input terminals of the receiver, at least when the difference q_s between the frequencies of the interfering oscillation and the carrier is not larger than q_a. With a larger frequency difference the interfering oscillations are inaudible.

In a frequency modulation receiver the situation is different. There the frequency detector converts a frequency modulated oscillation into an oscillation which also contains amplitude modulation, as was already explained in connection with fig. 9 in the article referred to in footnote [1]). With a frequency sweep sq_s the amplitude modulation depth becomes sq_s/q_1, while the frequency of this modulation amounts to q_s. The oscillation *c* in question contains, however, also amplitude modulation with a depth of modulation *s*, and the same frequency q_s. The instantaneous amplitude of the oscillation produced by the frequency detector is proportional to the instantaneous amplitude of the supplied oscillation. The amplitude modulation of the oscillation *c* thus also has an influence on the amplitude modulation of the oscillation which the frequency detector delivers. The voltage which the frequency detector delivers, as a function of the instantaneous frequency, is now not represented by the straight part *AB* of the characteristic (*fig. 3*), but by the line *CD* or the line *EF*, according as the frequency of the interfering oscillation *b* is lower or higher than the frequency of the carrier *a*, as may be deduced from the vector diagram of fig. 2. The resultant depth of modulation of the voltage delivered by the frequency detector is $s\ (1 \pm q_s/q_1)$. It therefore seems justified in this simple case simply to add together or to subtract from each other the interferences *s* and sq_s/q_1, which would be caused by the amplitude and frequency modulation respectively independently of each other. This is in general not permissible, as we shall explain later in this article. In order to obtain the great advantage of frequency modulation, namely the slight influence of interferences, the maximum frequency sweep *q* should be smaller than q_n; the angular frequency corresponding to the limit of audition (as will appear below) q_s/q_1 is thus always smaller than unity.

---

[2]) In places where it can cause no confusion we shall speak of frequency when angular frequency is meant, as in the article referred to in footnote [1]).

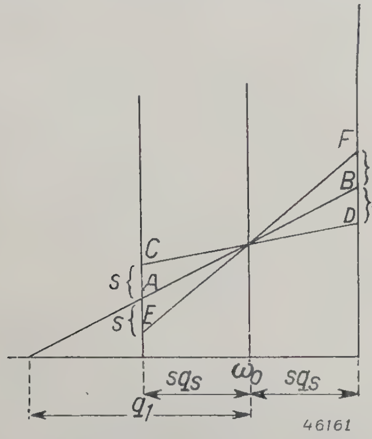


Fig. 3. Variation of the voltage furnished by the frequency detector when the signal is both amplitude and frequency-modulated. The line *AB* is the approximately straight part of the characteristic which upon extension intersects the frequency axis at a distance q_1 from ω_0.

The first term in $s\ (1 \pm q_s/q_1)$, caused by the amplitude modulation which the carrier has received from the interference, therefore has the greatest influence. The interference can now in general be diminished by taking care that this term does not occur. This can be achieved by introducing a limiter in front of the frequency detector, which provides that the amplitude of the oscillation which reaches the frequency detector is constant; then only the frequency modulation remains, so that the amplitude modulation depth of the oscillation which the frequency detector furnishes only amounts to sq_s/q. In the vector diagram of fig. 2 the action of the limiter is manifested in the reduction of the length of the vector *OQ* to a small amount, so that the end point of this limited vector moves along the circular arc *BC* around *O*.

Various methods may be followed for the limitation. *Fig. 4* shows a simple, but quite satisfactory circuit. Its action is based upon grid detection. With increasing amplitude of the A.C. voltage on the grid the negative grid bias becomes larger, whereupon the amplification of the valve decreases. With a suitable choice of the elements of the connections, the amplitude of the output voltage depends scarcely at all on the input voltage, within certain limits. *Fig. 5* shows the change taking place in the original signal *d* by the limiter. Since according to fig. 4 the latter contains resonance circuits, it will be clear that *d* is not simply cut off at *e* horizontally, but that the smooth curve *f*, here drawn with a dotted line, is the result of the limitation. If we assume that the limitation is ideal, thus that it furnishes a frequency-modulated oscillation with a constant amplitude, after detection a low-frequency signal occurs whose amplitude is proportional to sq_s/q_1. If in the absence of interference the signal is sinusoidally frequency-modulated with a frequency sweep $\Delta\omega$, the detector furnishes a desired low-frequency oscillation whose amplitude is proportional to $\Delta\omega/q_1$. The ratio of the

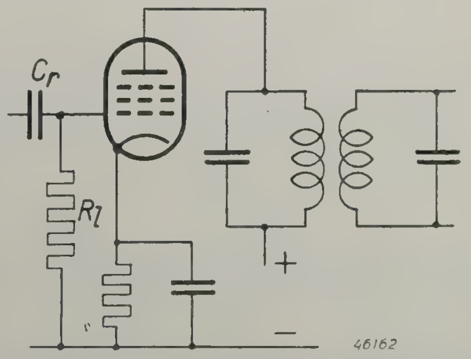


Fig. 4. Diagram showing the principle of an amplitude limiter for a frequency-modulation receiver. C_r grid condenser and R_l leakage resistance.

amplitude of the interference during a pause in the desired modulation to the amplitude of the desired signal upon modulation with a frequency swing $\Delta\omega$ is, therefore, $sq_s/\Delta\omega$. The frequency of this interference is q_s and it can therefore only be heard when $q_s < q_a$. With amplitude modulation this ratio was s. With frequency modulation the influence of the interference may therefore be made much smaller than with amplitude modulation, namely in the ratio $q_s/\Delta\omega$ when the amplitude is considered, *i.e.* in the ratio $(q_s/\Delta\omega)^2$ when the energy is considered. This should be done by taking $\Delta\omega$, the frequency sweep, of the desired modulation much larger than q_a, the highest modulation frequency to be transmitted (this is the modulation frequency corresponding to the limit of audition).

From these considerations the conclusion may be drawn that in a receiver for frequency modulation with limiter a simple sinusoidal interference with a small amplitude leads to a less audible interference than is the case in a receiver for amplitude modulation, when the frequency difference between interference and carrier is less than the angular frequency q_a corresponding to the limit of audition. Without a limiter, however, the audible interference may be larger than in the case of an amplitude-modulation receiver. Interferences with larger frequency differences than q_a are inaudible in both methods of modulation, provided they do not occur simultaneously with the desired modulation.

The case where a sinusoidal interference of small amplitude and the desired modulation are present at the same time.

We shall now examine what audible interference is caused by an oscillation with a small amplitude, whose frequency deviates only little from that of the carrier to be received, when the carrier is modulated with an audible frequency. For the sake of comparison we shall first consider the case where the desired signal is amplitude-modulated, while our receiver is intended for amplitude-modulated signals. We shall then consider the case where the desired signal is modulated in frequency, while our receiver is designed for frequency-modulated signals. A sinusoidal amplitude-modulated signal can be represented by $(1 + m \sin pt) \sin \omega_0 t$; the interfering oscillation by $s \sin (\omega_0 + q_s)t$. At small values of s the sum of these oscillations can approximately [3]) be represented by $(1 + m \sin pt + s \cos q_s t) \sin (\omega_0 t + s \sin q_s t)$. The detector in an amplitude-modulation receiver gives as low-frequency oscillation practically: $m \sin pt + s \cos q_s t$. The interference is now heard if $q_s < q_a$; if $q_s > q_a$ no interference can be heard. The amplitude of the interfering sound divided by the amplitude of the desired sound is in this case s/m; at 100 percent depth of modulation of the desired signal this ratio becomes s, just as in the case of the alternate occurrence of the interference and the desired signal.

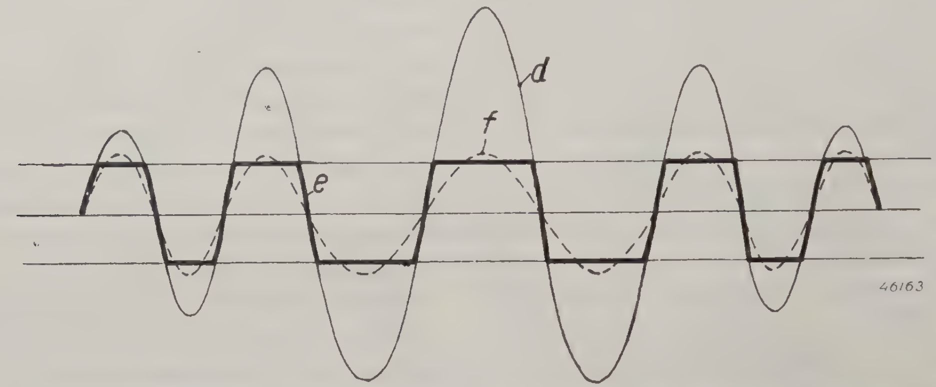


Fig. 5. If the amplitude limiter would only cut off the too high peaks of the amplitude de modulated signal d the heavier drawn line e would remain. In reality the amplitude limiter furnishes the dotted flowing curve f.

A sinusoidal frequency-modulated oscillation can be represented by $\sin (\omega_0 t + m \cos pt)$; the interfering oscillation is again $s \sin (\omega_0 + q_s)t$. At small values of s the sum of these oscillations can be approximately represented by:

$$\{1 + s \cos (m \cos pt - q_0 t)\} \cdot$$
$$\cdot \sin \{\omega_0 t + m \cos pt - s \sin (m \cos pt - q_0 t)\}.$$

We begin with the assumption that our frequency-modulation receiver contains a limiter, which is always the case in practice. This reduces the amplitude of the oscillation to a constant amount, whereby the oscillation which reaches the frequency detector is only proportional to:

---

[3]) Stricktly speaking this approximation is only permissible as long as $s < 1 - m$, which means that at all times the interference s must be small compared to the momentary amplitude of the desired signal; with great depth of modulation (*i.e.* when m is approximately unity) the formula becomes more complicated. See also T. Ned. Radiogen. 8, 340, Oct, 1940.

$$\sin \}\omega_0 t + m \cos pt - s \sin (m \cos pt - q_o t)\{$$

The instantaneous frequency of this is:

$$\omega_\mathrm{m} = \omega_0 - \Delta\omega \sin pt + s\,(q_\mathrm{s} + \Delta\omega \sin pt) \cdot \\ \cdot \cos\,(m \cos pt - q_\mathrm{s} t),$$

where $\Delta\omega = mp$. (*cf.* the article referred to in footnote [1]) represents the frequency sweep. The low-frequency signal which is applied to the loudspeaker is proportional to the difference between the carrier frequency and the instantaneous frequency, thus to:

$$\omega_0 - \omega_\mathrm{m} = \Delta\omega \sin pt - s\,(q_\mathrm{a} + \Delta\omega \sin pt) \cdot \\ \cdot \cos\,(m \cos pt - q_\mathrm{s} t).$$

The first term of this represents the desired signal, the second term the interference. This interference is an oscillation whose amplitude as well as whose frequency changes with the time. The instantaneous frequency of the interference, according to the definition given in the article referred to in footnote [1]), is $q_s + \Delta\omega \sin pt$. Since the preceding selective circuits do not pass the higher frequencies, in the case of the frequency detector only those values of q_s come into consideration which are smaller than q_a, while the maximum value of $\Delta\omega$ is a little smaller than q. The instantaneous frequency of the interference in the detected low-frequency signal thus varies between zero and $q_s + \Delta\omega$. But the velocity at which this instantaneous frequency varies is in general not always small compared with the instantaneous frequency itself; it is therefore of little use here to speak of the instantaneous frequency. The audible interference here has the character of a noise which, in the commonest practical case, namely when $\Delta\omega > q_a$, is only heard for part of the low-frequency cycle. If nevertheless, for the sake of simplicity, but then only in rough approximation, it is desired to continue to speak of instantaneous frequency, it may be said that the interference can only be heard when its instantaneous frequency lies between zero and q_a. We shall here omit a calculation of the ratio between the audible interference and the audible desired signal. In most practical cases, however, the ratio of the energy of the audible interference to the energy of the desired audible signal is smaller with frequency modulation than with amplitude modulation. On the other hand there are cases where with frequency modulation interference is indeed heard, but not with amplitude modulation. With amplitude modulation only those interfering oscillations produce an audible interference whose frequencies lie in the region of $\omega_0 - q_a$ to $\omega_0 + q_a$, while with frequency modulation this region extends from $\omega_0 - q_1$ to $\omega_0 + q_1$, where as a rule $q_1 > q_a$.

Interferences by noise

After having explained in the foregoing the influence of a single sinusoidal interference with a small amplitude, we shall now investigate how the receiver reacts to interferences of less simple nature such as often occur in practice. In the first place we shall consider the interferences due to noises which are caused by the thermal motions of the electrons in resistances and by the corpuscular structure of electricity (electrons) in amplifier valves. This noise can be conceived as an oscillation in which all frequencies occur in equal intensity. As in the foregoing chapter, we shall first study the effect of the noise during a pause in the modulation of the desired signal. For the sake of simplicity we consider the source of noise to be localized at the input terminals of the receiver, which is always permissible because all cases of noise can be reduced to that case. Let us now in the first place consider the case of an amplitude-modulation receiver. Due to the noise at the input terminals uniformly distributed over the whole frequency spectrum we then obtain a noise at the loudspeaker likewise uniformly distributed over all audible frequencies. The energy of the noise at the input terminals of the receiver in every arbitrary frequency band dq_s we call $s^2 dq_s$, when the energy of the desired carrier is set equal to unity. The energy of the noise in the loudspeaker in every frequency band $\mathrm{d}q_s$ is also equal to $s^2 \mathrm{d}q_s$ when the energy of the detected desired signal is again set equal to unity at 100 percent depth of modulation. In order to find the total effect of the noise this energy must be integrated over the whole audible frequency region, and this is:

$$\int_0^{q_\mathrm{a}} s^2 \mathrm{d}q_\mathrm{s} = s^2 q_\mathrm{a}.$$

when the intensity s of the noise is uniformly distributed over the whole audible frequency region. The ratio of the energy of the noise interference to that of the desired signal is therefore $s^2 q_a$.

In the case of a frequency modulation receiver the frequency spectrum of the noise as it reaches the loudspeaker is different from that at the input terminals. In the consideration of a single sinusoidal interference we have already noted that in the presence of a limiter the amplitude-modulation depth of the oscillation which the frequency detector furnishes during a pause in the modulation amounts to sq_s/q_1. The energy which is represented by this oscillation is then proportional to $s^2(q_s/q_1)^2$. The energy of the noise as it reaches the loudspeaker in the frequency band from q_s to $q_s + dq_s$ is proportional to $s^2(q_s/q_1)^2\, dq_s$. The energy of the noise in the loudspeaker for the whole audible frequency region is then:

$$\int_0^{q} s^2 \frac{{q_s}^2}{{q_1}^2}\, dq_s = \frac{1}{3}\, s^2 \frac{{q_a}^3}{{q_1}^2}.$$

The amplitude of a desired detected signal caused by a sinusoidal frequency-modulated oscillation with a frequency sweep $\Delta\omega$ with the same proportionality factor is $\Delta\omega/q_1$; the energy is thus $(\Delta\omega/q_1)^2$. The ratio of the energy of the noise interference to that of the desired signal is thus $^1/_3\ s^2 q_a^3/(\Delta\omega)^2$. This ratio in the case of frequency modulation is $^1/_3\ (q_a^2/\Delta\omega)^2$ times that in the case of amplitude modulation. Frequency modulation is in this respect therefore more advantageous than amplitude modulation if the frequency sweep $\Delta\omega$ is taken much larger than the highest audible frequency q_a to be transmitted, as was the case with a single sinusoidal interference.

We shall explain this result on the basis of a figure. The amplitude of the low-frequency signal after detection of a frequency-modulated oscillation is proportional to the frequency sweep. After detection the amplitude of the single sinusoidal interference in the absence of the modulation of the desired signal, divided by the amplitude of the desired signal with sinusoidal modulation, is equal to $sq_s/\Delta\omega$. In *fig. 6* OA represents the amplitude 1 and OB the maximum frequency sweep $\Delta\omega$ of the desired signal; BC the amplitude s as this was defined in the foregoing for an interference by noise, the line OC the amplitude modulation the amplitude increases linearly with increasing q_s; for $q_s = 0$ this amplitude is zero, for $q_s = q_a$ this amplitude is $sq_a/\Delta\omega = DE$ in fig. 6. Only the noise components whose frequency is smaller than q_a (OD in fig. 6) contribute to the audible interference. The energy of the interfering noise in a given frequency band dq_s is proportional to the square of the amplitude. The total audible noise energy is thus found by integrating the squares of the ordinates in fig. 6 from zero to q_a. When this integration is performed the already mentioned ratio $(^1/_3\ q_a/\Delta\omega)^2$ is again obtained for the energy of the noise with respect to that of the desired signal.

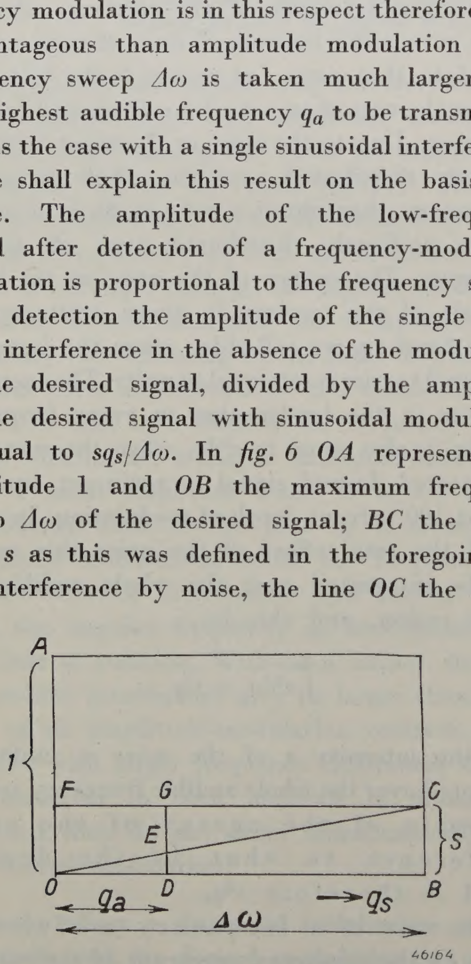


Fig. 6. The amplitude of the detected interference in the case of a frequency-modulation receiver as a function of the frequency difference q_s between the desired carrier and the interfering signal is represented by the straight line OC through the origin. In the case of an amplitude modulation receiver the horizontal line FG gives the corresponding relation.

Since the influence of the noise in frequency modulation can be made small by choosing the frequency sweep $\Delta\omega$ much larger than the limiting frequency q_a of the audible region, it might be thought that by choosing $\Delta\omega$ larger and larger the noise could be diminished to an unlimited extent. This, however, is incorrect, since the frequency sweep $\Delta\omega$ cannot be increased directly above a definite limit. We shall, however, return later to this question.

In the foregoing discussion it was always assumed that a limiter was present in the receiver in front of the frequency detector; moreover, the influence of the noise was only studied during a pause in the desired modulation. If, however, the limiter is absent and if the desired and undesired modulations are present simultaneously, one finds for the relation of the quotient of noise energy and the energy of the desired signal with frequency modulation to that with amplitude modulation:

$$\frac{1}{3}\left(\frac{q_a}{\Delta\omega}\right)^2 + \left(\frac{q_1}{\Delta\omega}\right)^2 + \frac{1}{2}.$$

which expression we shall not here derive. The first of these three terms is the same as what we found above for the case where a limiter was present and is therefore due to the frequency modulation caused by the noise. The second term is at least equal to unity and represents the interference contribution caused by the amplitude modulation which the unmodulated carrier receives as a result of the noise. The origin of the third term is not so easy to indicate, since it only occurs when interference and desired frequency modulation are present simultaneously, so that we may call it the interaction term. It may be remarked in conclusion that the interference energy due to noise in the case of frequency modulation without limiter is more than $1^1/_2$ times as large as in the case of amplitude modulation. If, however, there is a limiter present then also during the tude of the detected interference as a function of the frequency difference q_s between the desired carrier and the interfering signal in the case of a frequency modulation receiver. With amplitude modulation all audible noise components take on the same amplitude, namely *OF;* with frequency presence of the desired modulation the above mentioned relation is the same as during a pause in the desired modulation, *i.e.* only $^1/_3\ (q_0/\Delta\omega)^2$, which with a sufficiently large value of the frequency sweep $\Delta\omega$ is small compared with unity.

The effect of the limiter thus consists in the second and third terms, which amount respectively to at least 1 and $^1/_2$, being eliminated, so that only the first term is retained. It is therefore advisable to make the first term as small as possible, as has just been explained.

In practice, however, a limiter is never perfect; some influence of the amplitude modulation always remains. If, however, a pushpull detector is used, as discussed in the article referred to in footnote [1]) where the output voltages of the two diodes are opposite, then even when there is no limiter present the influence of the amplitude modulation is eliminated in as far as the second term of the expression just discussed disappears, but the third term, the reciprocal action term, then remains. The best result is therefore obtained by using a limiter and a push-pull detector together. In that case the interference energy found experimentally is indeed found to correspond to the small first term of our expression.

Interferences with large amplitudes

Until now it has been assumed that the amplitude of the interferences due to noise or other causes was small compared with the amplitude of the desired signal. If, however, this is not the case, the decrease of the interference with frequency modulation compared with that with amplitude modulation is not so great. If the amplitude of the interference is larger than the amplitude of the desired signal, then the interference is not drowned out by the signal, but the signal by the interference. As long as the interference is smaller than about half the desired signal, frequency modulation offers advantages over amplitude modulation, although to a smaller degree than would correspond to the expression given for that: $(q_s/\Delta\omega)^2$ for discrete interferences and $^1/_3\ (q_a/\Delta\omega)^2$ for noise. If the average amplitude s of the noise is smaller than about 1/10 of that of the desired signal, the interference is practically not larger than the expression gives. If for example we choose the frequency sweep $\Delta\omega = 5q_a$, the ratio of the interference energy with frequency modulation to that with amplitude modulation is 1/75, thus in our case the average amplitude of the noise: $\sqrt{1/75} \cdot 1/10 = 1/86$ of the amplitude of the desired signal, which still makes an acceptable reception possible. With amplitude modulation, however, with a noise with an average amplitude equal to 1/10 of the amplitude of the desired signal, satisfactory reception of broadcasting is out of the question. It has been found in practice that a voltage of about 10 μV on the input terminals of a well-constructed frequency modulation receiver is sufficient for satisfactory noise-free reception in the absence of other interferences, while with amplitude modulation about 1000 μV are necessary under the same conditions.

The amplitude of the interferences at the limiter is the decisive factor in this respect. This amplitude depends upon the frequency region passed by the band filters; the width of this region depends upon the maximum frequency sweep for which the receiver is built, because the high-frequency and intermediate-frequency band filters must pass a frequency band of at least $2\Delta\omega$. The average amplitude of the noise is proportional to the square root of the band width. The larger the frequency sweep $\Delta\omega$, therefore, the larger the noise amplitude at the limiter. But, as was already noted, the characteristic advantages of frequency modulation disappear if the noise amplitude at the limiter is no longer smaller than about half the amplitude of the desired signal. The frequency sweep cannot therefore without danger be increased indefinitely. If one assumes 15 000 cycles per second to be the highest number of oscillations per second to be observed, which therefore it is desired to receive from the desired modulation and which the low-frequency part of the receiver including the loudspeaker must reproduce, the practically most favourable value for the frequency sweep is about 75 000 cycles per second; thus calculated as angular frequency: $\Delta\omega = 2\,\pi \cdot 75\,000$.

Interferences from electrical apparatus and atmospherics

The influence of these interferences, which usually consist of several short impulses, cannot be accurately given in a general form because of their irregular character. With equal effective values, the maximum values of these interferences are in general larger than those of interferences due to noise. If the maximum value of the interferences reaches the level of the desired signal, the interference will be considerable during these maxima, *i.e.*, not smaller than with amplitude modulation, while during the time that the momentary value of these interferences is small, little or no disturbance is caused by them. The character of these interferences is thus quite different with frequency modulation than with amplitude modulation, in that with not too violent interferences frequency modulation with a larger frequency sweep is still more advantageous than amplitude modulation.

Interference by other transmitters

With respect to interferences by other transmitters frequency modulation is in general less subject to disturbances than amplitude modulation. An amplitude-modulated interfering transmitter with a frequency lying in the region that is occupied by the instantaneous frequency of a desired frequency-modulated transmitter behaves in this respect in the same way as a single sinusoidal interference. Thus if the interfering signal is much smaller than the desired signal it will cause little trouble. A frequency-modulated interfering transmitter which gives a smaller voltage in the receiver than the desired transmitter, and which occupies a frequency region coinciding partially or wholly with that of the desired frequency-modulated transmitter, will give little interference, in the first place for the same reason as in the case of other interferences and in the second place because in a first approximation an interference is only heard when the difference between the "instantaneous" frequencies of the two transmitters lies in the audible region, which will only be the case during part of the time, especially if the frequency sweep is large. If, however, the interfering frequency-modulated signal is not much smaller than the desired frequency-modulated signal, it is necessary, for interference-free reception, that the regions of the instantaneous frequencies of the two transmitters should not overlap. The frequency distance between the carriers must therefore in that case amount to at least twice the frequency sweep.

Conclusions

We have explained that the use of frequency modulation in radio broadcasting may under certain conditions offer advantages over amplitude modulation. These advantages consist mainly of less trouble with noise and other interferences, provided the amplitude of these interferences is smaller than the amplitude of the desired signal. In order actually to obtain these advantages it is necessary that an amplitude limiter should be introduced into the receiver, whilst, moreover, a push-pull detector is desirable. The frequency swing must be considerably larger than the highest low frequencies to be transmitted. A favourable practical value for the frequency sweep for broadcasting is 75 000 cycles per second. With this large frequency sweep the region practically occupied in the frequency spectrum is almost independent of the modulation frequency. It is therefore possible that the audio-frequency region can be extended to the auditory limit, for example, 15 000 cycles per second, without increasing the region occupied by the frequency spectrum, whereby the quality of reception can be improved to the utmost. If it is indeed desired to profit by this great advantage it is necessary that the transmitter (from microphone to aerial) as well as the receiver (from aerial to and including loudspeaker) should be designed for that purpose. If, however, this extension of the frequency region is actually to produce an improvement in quality, it is in addition required that the distortion in transmitter and receiver should be unusually small.

In the case of radio broadcasting with amplitude modulation it is customary to diminish the large differences in intensity which are indispensable for enjoyable reproduction of music[4]). This is necessary in that case, since otherwise the soft passages would be drowned out by the noise and the strong passages would be very much distorted. In the case of frequency modulation, however, this is not necessary since in practically the whole region in which the transmitter can be received free of interference no noise can be heard even during the softest passages. With frequency modulation, therefore, the broadcasting company is able to attain a quality of reproduction of music in which no violence has been done to its dynamics. Due to the large frequency sweep frequency modulation in the manner described can only be employed on short waves ($\lambda < 10$ m); for longer waves amplitude modulation is therefore to be preferred. Short waves can only be received in a limited region since they are not reflected by the ionosphere. Because of this the same carrier frequency can be repeated at a certain distance (for instance 300 km) without difficulty. Since the amplitude ratio of interfering and desired transmitters in the case of frequency modulation does not need to be so small for freedom from interference as with amplitude modulation, the aera of the earth's surface where two transmitters with the same carrier frequency interfere with each other is moreover much smaller.

---

[4]) See for instance: V. Cohen Henriquez, Philips techn. Rev. 3, 204, 1938.